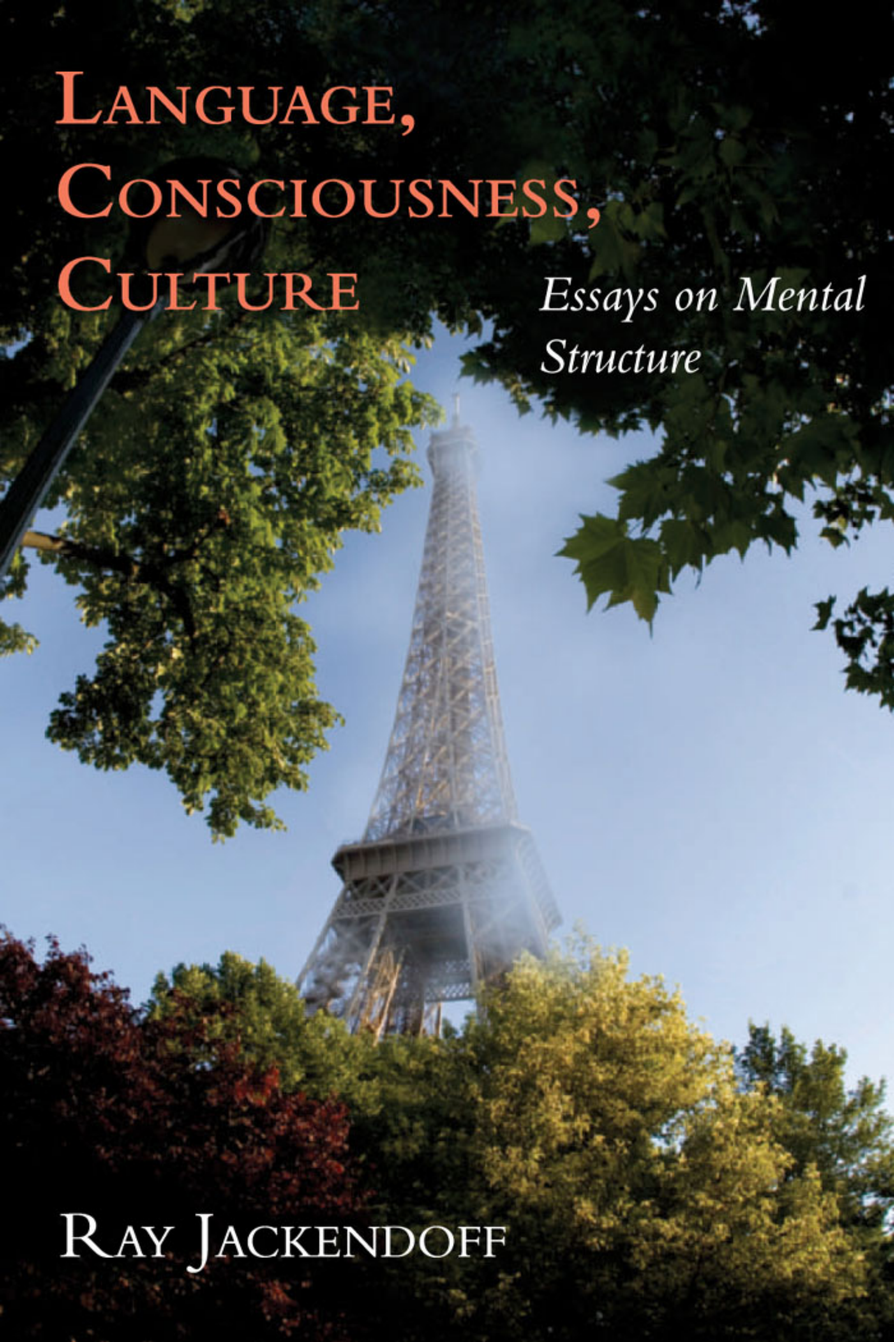
Language, Consciousness, Culture

Essays on Mental Structure



Ray Jackendoff

Language, Consciousness, Culture

The Jean Nicod Lectures
François Recanati, editor

The Elm and the Expert: Mentalese and Its Semantics, Jerry A. Fodor (1994)
Naturalizing the Mind, Fred Dretske (1995)
Strong Feelings: Emotion, Addiction, and Human Behavior, Jon Elster (1999)
Knowledge, Possibility, and Consciousness, John Perry (2001)
Varieties of Meaning, Ruth Garrett Millikan (2004)
Sweet Dreams: Philosophical Obstacles to a Science of Consciousness, Daniel C. Dennett (2005)
Language, Consciousness, Culture: Essays on Mental Structure, Ray Jackendoff (2007)

Language, Consciousness, Culture

Essays on Mental Structure

Ray Jackendoff

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please e-mail special_sales@mitpress.mit.edu or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong, and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Jackendoff, Ray, 1945–
Language, consciousness, culture : essays on mental structure / by Ray Jackendoff.
p. cm. — (The Jean Nicod lectures)
"A Bradford Book."
Includes bibliographical references and index.
ISBN-13: 978-0-262-10119-6 (hardcover : alk. paper)
1. Psycholinguistics. 2. Consciousness. 3. Mind and body. 4. Language and culture. I. Title.
P37.J327 2007
401.9—dc22 2006030131

10 9 8 7 6 5 4 3 2 1

For Hildy

Contents

Series Foreword

The Jean Nicod Lectures are delivered annually in Paris by a leading philosopher of mind or philosophically oriented cognitive scientist. The 1993 inaugural lectures marked the centenary of the birth of the French philosopher and logician Jean Nicod (1893–1931). The lectures are sponsored by the Centre National de la Recherche Scientifique (CNRS) and are organized in cooperation with the Fondation Maison des Sciences de l'Homme (MSH Foundation). The series hosts the texts of the lectures or the monographs they inspire.

Jacques Bouveresse, President of the Jean Nicod Committee
François Recanati, Secretary of the Jean Nicod Committee and
Editor of the Series

Preface

This book grew out of a most gracious invitation by the Institut Jean Nicod to present the Jean Nicod Lectures in Cognitive Philosophy in Paris in the spring of 2003. Given the broadly interdisciplinary nature of the Institut, I thought it would be fun to offer a fairly wide-ranging series of lectures in a somewhat speculative vein. And I seem to have succeeded: the audiences at the lectures were warm and engaged, and I was delighted with the lively and useful discussion.

The overarching topic of the book is an exploration of the mental structures involved in a variety of cognitive domains: language, consciousness, complex action, theory of mind, and social/cultural cognition. I use the term "mental structures" rather than the more traditional "mental representations" for reasons discussed in chapter 1. The notion of rigorously investigating mental structure is familiar from linguistics, but has had little currency in the rest of cognitive science. Part of my goal is to demonstrate that even in this age of neuroscience, where the ultimate goal is understanding the structure of the brain, there is still a lot to be learned by attempting to describe the more abstract level of mental structure, where issues of combinatoriality can be addressed in a fashion as yet impossible in neural terms.

Part I, the first five chapters, is an augmented version of the Jean Nicod Lectures. Chapter 1 presents an account of what I mean by mental structure as a formal system, how it is related to brain structure as studied by neuroscience, and how this relation affects issues such as processing, learning, and modularity. Chapter 2 summarizes the central arguments of my books *Foundations of Language* and *Simpler Syntax* (the latter in collaboration with Peter Culicover). It considers why linguistics has become intellectually isolated from the other cognitive sciences (without most linguists noticing or caring). The conclusion is that although there are undoubted sociological and historical reasons for this situation, there are also scientific reasons, growing out of fundamental assumptions about the architecture of language, inherited without question from the early days of generative grammar. The chapter sketches the alternative of a *parallel architecture*, a conception of the overall structure of language that is more in tune with contemporary empirical evidence about the relation among syntax, phonology, and semantics than is the predominant "syntactocentric" approach. The chapter goes on to demonstrate that the parallel architecture is superior to the classical architecture in the approach it affords to language processing, to the overall organization of the brain, and to the evolution of the language capacity.

With a theory of language in hand that answers to larger issues in cognitive neuroscience, we are poised to extend the fundamental questions of mental structure beyond the language capacity. Chapter 3 updates the inquiry into consciousness undertaken in my 1987 book *Consciousness and the Computational Mind.* It poses a counterpart of the neuroscientific question of the Neural Correlates of Consciousness: what are the *mental structures* that are most closely correlated with the character of experience? Posing this question in terms of language—the mental faculty whose structures we understand best—sharpens the criteria for a satisfactory theory of consciousness. It proves easy to reveal fundamental flaws in most of the influential theories in the literature—theories that are confined to visual perception and hardly address other modalities of experience. In particular, examining consciousness in terms of mental rather than (or in addition to) neural structures makes it possible to characterize the phenomenology in much more precise terms than is possible in other approaches, and it allows us to state clearly the issues involved in the relations among language, thought, and awareness.

Chapter 4 was something of a surprise to me. Since topics such as intention, obligation, and social norms, to be studied in the rest of the book, are conditions not on beliefs but on actions, I felt it would be important to understand something about the structure of action. My explorations led to simpler and simpler actions, while still revealing surprising complexity—much of which had been established previously, especially by researchers in robotics. What is novel here is the discovery of significant parallels between the capacity for complex action and the capacity for language production. One outcome is a new take on what is special and what is not special about the language capacity, a hot topic in the current debates on the biological and evolutionary foundations of language.

Chapter 5 deals with a domain that has attracted me for some years: the mental capacities involved in an individual's grasp of society and culture. I have written before on this topic (in *Languages of the Mind* and *Patterns in the Mind*), but this is my most detailed exposition to date. The essential idea is that, like language, culture is learned by individuals; but, like language, it is probably learned by virtue of an innate basis with evolutionary antecedents. This position, rather far out at the time I began exploring it, is now very much in the mainstream among evolutionary psychologists and cognitive anthropologists. What seems still to be new here is the focus on the formal organization of the abstract concepts involved in social/cultural knowledge such as group membership, rights and obligations, values, and reputations, as well as the often peculiar inferences that invoke these concepts. Thus this focus offers a prospect for a more rigorous investigation of what it takes to be a socially interacting human being. In particular, it helps to distinguish social competence per se from such related issues as theory of mind, and to open up the scope of investigation to a far broader range of phenomena, some of which reappear in later chapters.

Part II interlocks with part I. It takes up the challenge posed by chapter 5, developing formal analyses of concepts involved in social cognition and theory of mind. The inquiry is conducted within the overall framework of Conceptual Semantics developed in my books *Semantics and Cognition*, *Semantic Structures*, and *Foundations of Language*. Conceptual Semantics, unlike influential approaches arising from the philosophical tradition, is intended as a theory of meaning *as it is instantiated in the mind*; it thus has rich interactions with cognitive neuroscience and evolutionary psychology. Conceptual Semantics and related approaches, especially within Cognitive Grammar, have been extraordinarily successful in stimulating research in spatial cognition. Chapters 6–11 break new ground in moving to the social domain.

Chapter 6 works out an account of perception verbs such as *look* and *see*, showing that *look* is in a sense "objective," but *see* is "subjective" and takes into account theory of mind. In addition, it shows how this semantic analysis reflects on the general problem of linking the semantic arguments of verbs to syntactic positions such as subject and object. Chapter 7 extends the machinery to affective/evaluative predicates such as *interesting* and *fascinated*. Here again, one focus is the distinction between ostensibly objective evaluations (e.g. *This topic is interesting*) and subjective evaluations (e.g. *This topic interests me*).

Chapter 8 is concerned with intending and volitional action, and with the relation of intending to that quintessential propositional attitude, believing. Like chapters 6 and 7, it shows how the syntactic patterns associated with verbs that express attitudes are partly a consequence of the verbs' semantics. The chapter also develops a formal characterization of Dennett's notion of the "intentional stance" and its relation to theory of mind, as well as a formal account of joint action and joint intention, crucial to an account of cooperation.

The topic of chapter 9 is values of all sorts: the value of an action or an object to an individual, normative values of actions (morality, etiquette, etc.), the moral worth of an individual, and the esteem in which an individual is held. Like the evaluative predicates in chapter 8, values come in objective and subjective flavors, and chapter 9 explores the consequences of this distinction, as well as the peculiar logic that links all the different sorts of value and helps guide action. Chapter 10 applies this logic of values to develop a conceptual account of fairness and reciprocation. In particular, it draws a distinction rarely recognized in the literature between freely undertaken *reciprocation* (which includes reciprocal altruism) and agreed-upon *exchange*, a joint undertaking with quite a different logic. It also shows that reciprocation has strong parallels in the use of displays of esteem or respect, a category of social action rarely dealt with in the literature on norms and morality. Chapter 11 turns to rights and obligations, which form an indispensable basis for social organization in every culture: they lie behind promises, contracts, marriages, laws, and privileges of authority, and their justification and enforcement create one of the principal motivations for both government and religion. Chapter 12 wraps matters up with a return to the larger issue of what makes humans special.

The discussion in chapters 6–12 veers freely between strict linguistic semantics and more general concerns in consciousness, theory of mind, theory of action, social cognition, and moral theory. Although I have tried to make the formal treatment reader-friendly, it is still probably a challenge. I urge readers nevertheless to stay the course, because issues of interest to a broader spectrum of readers in cognitive science tend to emerge at unexpected places in the formal treatment. These generalizations could not have been discovered in the absence of a suitable formal framework.

For the most part, the chapters are independent of each other, although chapter 1 is a useful introduction to any of them, and chapter 5 is a useful introduction to part II. On the other hand, connections among the chapters keep cropping up, especially in part II, and not by accident. In particular, chapter 7 builds on chapter 6, and chapter 10 builds on chapter 9. Still, to make this into a fully unified volume would take many more years of work. I see the book, then, as offering preliminary snapshots of a territory that I find fascinating, and whose value I hope to persuade my colleagues to appreciate.

Acknowledgments

My deepest appreciation goes to the members of the Institut Jean Nicod for the honor and the unparalleled opportunity of presenting the Jean Nicod Lectures. In particular, I must thank François Recanati, who coordinated my visit, along with his active coconspirators Dan Sperber, Pierre Jacob, Vanessa Nurock, and Nausicaa Pouscoulos.

In addition to the Jean Nicod community, I am grateful to Dan Dennett, Marc Hauser, Tecumseh Fitch, Marcel Kinsbourne, Deb Roy, James Pustejovsky, Janet McIntosh, Barbara Landau, Wally Chafe, Judy Kegl, Anne Senghas, Laura Salmon, Robert Van Valin, Marion Smiley, Steve Umans, Stuart Shapiro, Martha Palmer, Jay Keyser, Larry Solan, Carlos Otero, Storrs McCall, Al Bregman, Paul Bloom, Jay Conison, Gordon Bower, my late father Nathaniel Jackendoff, my brother Harry Jackendoff, my daughter Beth Jackendoff, and my wife Hildy Dvorak for vast amounts of valuable input on this material. I must also thank the members of my spring 2005 course "Cognition of Culture and Society," taught jointly with Janet McIntosh at Brandeis University; their questions and comments and above all their rapt engagement with these issues were invaluable in showing me how to present the material in chapters 5, 9, and 10.

For material comforts as well as intellectual stimulation, I thank Marc Hauser, who provided me an office in the Psychology Department at Harvard during my sabbatical year 2002–2003, while I was working up the Jean Nicod Lectures. The support of NIH grant DC 03660 (Edgar Zurif and David Swinney, principal investigators) helped a lot too.

Earlier versions of some of these chapters (much earlier versions, in some cases) have been published elsewhere and appear here with the gracious permission of their respective publishers:

Chapter 2: Précis of *Foundations of language: Brain, grammar, meaning, evolution. Behavioral and Brain Sciences* 26, 651–665. Cambridge: Cambridge University Press, 2003.

Chapter 8: The conceptual structure of intending and volitional action. In Héctor Campos and Paula Kempchinsky, eds., *Evolution and revolution in linguistic theory: Studies in honor of Carlos P. Otero*, 198–227. Washington, D.C.: Georgetown University Press, 1995.

Chapter 11: The natural logic of rights and obligations. In Ray Jackendoff, Paul Bloom, and Karen Wynn, eds., *Language, logic, and concepts: Essays in memory of John Macnamara*, 67–95. Cambridge, Mass.: MIT Press, 1999.

Figure 1.1 is borrowed from my *Foundations of Language* (2002a) and appears here by courtesy of Oxford University Press.

Finally, profoundest thanks go to my wife Hildy, who made the stay in Paris for the Nicod Lectures a true joy, and who, come to think of it, has that effect on everything in my life.

Language, Consciousness, Culture

PART I

The Nicod Lectures

Chapter 1

Mental Structure

1.1 Locating the Study of Mental Structure in Cognitive Neuroscience

This book is concerned with exploring human nature in terms of the mental structures that play a role in constituting human experience and human behavior. In order to explain what I mean by "mental structure," it is useful to situate the term within the more general enterprise of cognitive neuroscience.

The leading question of cognitive neuroscience is how the brain works, such that it supports or generates cognition—where by "cognition" I mean an organism's understanding or grasp of the world, and its ability to formulate and execute actions in the world. The neuroscience part of the enterprise includes the study of the physical structure and activity of the brain at all scales, from the inner workings of neurons to the overall organization of brain areas. The cognitive part includes characterizing the functional or computational character of mental activity, as well as the organism's phenomenology—how the organism experiences the world. I will use the term *brain* in the customary way to describe the physical body part which accomplishes cognition, and which is the proper domain of neuroscience. I will use the term *mind* to denote the brain seen from the point of view of its functional or computational aspect, and *mind/brain* when I wish to be neutral between the two.

An important goal of the enterprise is to figure out how the functional domain is instantiated in the neural domain—to use a now somewhat outdated analogy, how the brain's software runs on the hardware—and also to figure out how the neural and computational structures support conscious experience. At the moment, this goal seems far off. We know many details of how brain function is localized and many details of how individual neurons and small clusters of neurons function. But I think it is likely to be a long time before we understand how the neurons actually accomplish anything as complex as, say, language perception or the storage of vocabulary—in detail or even in principle. So the flood of recent advances in understanding the brain by no means undermines studies of the mind. Part of the burden of this book is to emphasize the value of investigating cognition in terms of mental structure.

Cutting across this dimension of the enterprise are developmental questions, at two scales. First, at the scale of the individual: how do the brain, mental functioning, and phenomenology develop in the individual from conception to death? And second, at the scale of evolution: how do characteristics of the species develop over evolutionary time under the pressures of natural selection? The latter question adds to the mix the fascinating issue of interspecies comparison.

Cutting across both these dimensions is how the functions of the mind/brain divide into capacities or domains or modules or faculties, whatever you wish to call them. On one hand, there is a "vertical" division more or less by subject matter: vision, audition, proprioception (the sense of body position and movement), motor control, language, and so forth. And on the other hand, cutting across *this* is a "horizontal" division into, on one hand, the study of mental structure, and on the other, the kinds of machinery that process mental structures, such as working memory, long-term memory, attention, and learning, all of which are involved in each of the "vertical" capacities. Table 1.1 sums up all the dimensions of the inquiry.

Of course, we often study an individual cell in this four-dimensional matrix as though it were isolated—say, the brain localization of some aspect of visual working memory. However, we should understand that the essence of the enterprise lies in characterizing the interaction of these systems.

It is my impression that of all the cognitive sciences, only linguistics has systematically and explicitly investigated the content of mental structures that underlie a human capacity. The rest of cognitive neuroscience has for the most part made do with relatively rudimentary notions of mental structure, exploring more intensely issues of neural localization and/or the "horizontal" capacities of working memory, attention, learning, and the like. Three exceptions: Marr 1982 is the inception of a detailed study of the mental structures involved in vision (with Biederman 1987 as a related endeavor); this style of investigation has receded since Marr's death. Lerdahl and Jackendoff 1983 applies the approach of linguistic theory to music cognition. Finally, chapter 4 compares language with the capacity for complex action.

Table 1.1
Ways of studying the mind/brain

Dimension 1
Neuroscience (brain) vs.
Cognitive science (mind/functional properties) vs.
Behavior and phenomenology
—*plus* relations among the three
Dimension 2
Steady state vs.
Individual development vs.
Evolutionary development
Dimension 3 ("Vertical" capacities or "modules")
Vision vs.
Language vs.
Motor control vs.
Abstract thought vs.
. . .
Dimension 4 ("Horizontal" division; applies to all "vertical" capacities)
Data structures (mental structures) vs.
Processing capacities
Working memory vs.
Long-term memory vs.
Attention vs.
Learning

1.2 Mental "Structure" versus Mental "Representation"

Since the early days of cognitive science, the term of art for the computational structures in terms of which the mind operates has been "mental representations" or "symbolic representations." The subtitle of this book deliberately substitutes "mental structures"; let me explain why. The structures that a linguist writes on the page, say syntactic trees, are intended as representations of what is in the mind. However, I would maintain that what is in the mind is best not thought of as a representation or a symbol of anything. The reason is that the words "representation" and "symbol" imply an interpreter or perceiver: it is not just that *this* represents or symbolizes *that*, but implicitly that *this* represents or symbolizes *that* ***to so-and-so***. But a person in whose mind syntactic structures reside does not perceive them; rather, the person perceives a linguistic utterance by virtue of having these structures in his or her mind. The only thing that "perceives" syntactic structures is the faculties of mind that process and store syntactic structures, and in fact the term "perceive" is itself suspect in this context.

If we are to take seriously the relation between mind and brain, this is the only possible view of mental structures. The neurons deep inside the brain that are responsible for cognition have no privileged access to the "real world"; they interact only with other neurons. Contact with the "real world" is established only through long chains of connection leading eventually to sensory and motor neurons. If this is the hardware on which mental capacities "run," then mental capacities too are necessarily limited in their contact with the "real world." They are sensitive to the outside environment only insofar as they are connected through functional (or computational) links to the sensory and motor capacities.

In short, I wish to reject all talk of the "intentionality of mental representations," the idea that mental structures are "about" the world in some direct sense. This goes against the grain of much influential philosophy of cognitive science (e.g. Searle 1980; Fodor 1987).[1] The reader is free to understand such rejection in either of two ways. The weaker stance is methodological: even if mental structures are ultimately connected directly to the world by intentionality, there remains the empirical enterprise of characterizing them for their own sake. Taking this stance, we are choosing to study mental structures as a kind of "engineering," temporarily leaving philosophical concerns behind.

The stronger stance is to take the rejection of intentionality as principled—to claim that once the mental structures are properly characterized, there will be no need for a supervenient intentionality. Such a stance fits far more comfortably with the neuroscience. On the other hand, it depends on a promissory note to the effect that *someday* all the problems associated with intentionality will be worked out. But of course we adopt such promissory notes all the time in science. In particular, any sort of materialist philosophy of mind (i.e. any sort of modern cognitive science) takes for granted the promissory note that someday we will be able to relate all mental processes to brain processes.

[1] For extended discussion of why I reject intentionality, see Jackendoff 1987, chap. 7; 1992a, chap. 8; 2002a, chaps. 9, 10. Some of the more confrontational commentaries on Jackendoff 2002a (e.g. Adams 2003; Higginbotham 2003; Gross 2005; Rey 2006) reflect the degree to which intentionality is still taken as a sine qua non of theories of mental representation.

For the working scientist, the choice between the methodological and the principled stance rarely affects one's work one way or the other. As far as I can see, the main thing that cripples inquiry is to proclaim that without an account of intentionality, all research on mental function is pointless, and to demand that intentionality be explained before any further work proceeds.

1.3 The Mental Structures of a Simple Sentence

This section presents a very elementary example of linguistic structures as linguists understand them; the next section briefly discusses the issues that such structures raise for neuroscience. Section 1.5 sketches an overall view of the character of the mind in these terms.

So consider someone saying an absolutely simple sentence such as *The little star's beside a big star*. This is quite likely a sentence the speaker has never uttered or heard before. The speaker has constructed it to suit some present communicative context, using elements from his or her long-term memory, in particular the words and the means of putting them together into sentences (the latter often called "rules of grammar"). Linguistic theory is primarily concerned with how words and the principles for combining them are to be characterized *functionally*—as mental data structures, so to speak.

Figure 1.1 (pp. 8–9) shows some of the more prominent aspects of the structure of the sentence *The little star's beside a big star*. These are aspects on which there is substantial agreement among linguists, whatever their creed (Chomskyan or not); there are many disagreements about what further complexity there might be, but there is at least this much. Let me give a brief tour of this structure. (There is more detail in chapter 2, and especially in Jackendoff 2002a, chaps. 1 and 5.)

The upper part of the figure works out the phonological (or sound) structure of the sentence. The basic pronunciation of the sentence appears on the line labeled "segmental structure"; each of the symbols in this line stands for a speech sound. There is substantial agreement that segmental structure is more articulated than this: each speech sound is actually a composite of *phonological distinctive features*. Figure 1.2 (p. 10) shows the decomposition of this level for just the word *star*; you can imagine extending this analysis to the rest of the sentence. The distinctive features capture the dimensions of variation among speech sounds, for instance the position of the tongue, jaw, lips, and velum, and the presence or absence of vocal cord vibration.

Phonological structure

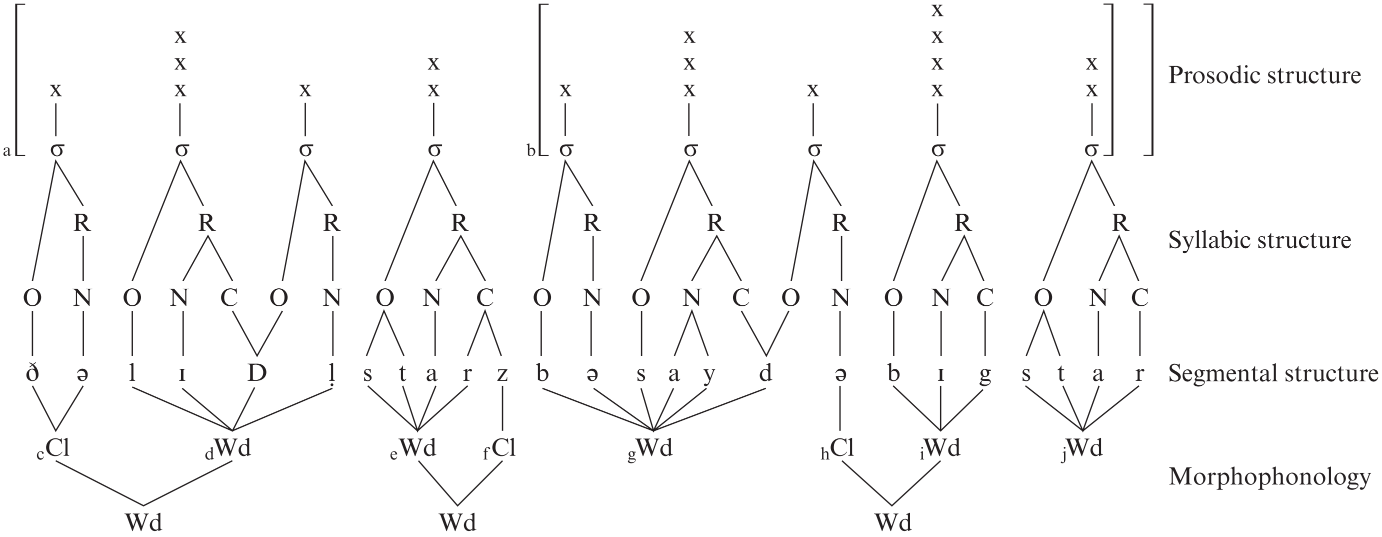


Figure 1.1
Structure of *The little star's beside a big star*

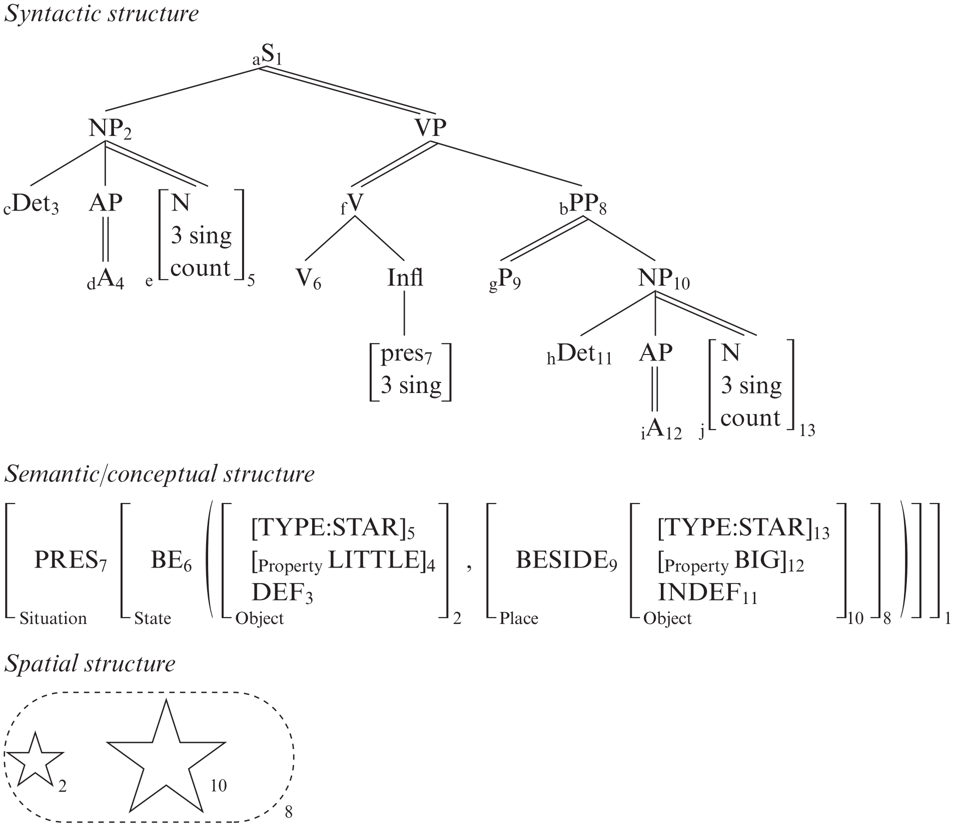

Syntactic structure
$_a S_1$
NP_2
VP
$_c Det_3$
AP
$_d A_4$
$_e$ [N 3 sing count]$_5$
$_f V$
V_6
Infl
[pres$_7$ 3 sing]
$_b PP_8$
$_g P_9$
NP_{10}
$_h Det_{11}$
AP
$_i A_{12}$
$_j$ [N 3 sing count]$_{13}$
Semantic/conceptual structure
$\left[_{\text{Situation}}\ \text{PRES}_7 \left[_{\text{State}}\ \text{BE}_6 \left(\left[_{\text{Object}} \begin{array}{l}[\text{TYPE:STAR}]_5\\ [_{\text{Property}}\ \text{LITTLE}]_4\\ \text{DEF}_3\end{array}\right]_2 , \left[_{\text{Place}}\ \text{BESIDE}_9 \left[_{\text{Object}} \begin{array}{l}[\text{TYPE:STAR}]_{13}\\ [_{\text{Property}}\ \text{BIG}]_{12}\\ \text{INDEF}_{11}\end{array}\right]_{10}\right]_8 \right)\right]\right]_1$
Spatial structure
2
10
8

Figure 1.1
(continued)

$$\begin{bmatrix} +\text{consonantal} \\ -\text{vocalic} \\ -\text{sonorant} \\ -\text{nasal} \\ +\text{continuant} \\ -\text{voiced} \\ +\text{anterior} \\ +\text{coronal} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ -\text{vocalic} \\ -\text{sonorant} \\ -\text{nasal} \\ -\text{continuant} \\ -\text{voiced} \\ +\text{anterior} \\ +\text{coronal} \end{bmatrix} \begin{bmatrix} -\text{consonantal} \\ +\text{vocalic} \\ +\text{sonorant} \\ -\text{nasal} \\ -\text{high} \\ +\text{low} \\ -\text{front} \\ -\text{round} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ +\text{vocalic} \\ +\text{sonorant} \\ -\text{nasal} \\ +\text{continuant} \\ +\text{voiced} \\ -\text{anterior} \\ +\text{coronal} \end{bmatrix}$$

Figure 1.2
Detail of segmental structure of *star*

Next let's return to figure 1.1. Above the segmental structure is a sequence of little tree structures that show how the speech sounds are collected into syllables (notated as σ in the trees). Each syllable contains a syllabic nucleus (N) and sometimes an onset (O) and coda (C). The nucleus and coda together form the rhyme (R), the part of the syllable that is used in determining rhyme, and also the part of the syllable that is relevant for determining stress.

Above the syllabic structure is a *metrical grid* of *x*s that marks the relative stress of the syllables in the sentence: more *x*s above a syllable indicate more stress. Thus the word *the* is relatively unstressed, and the word *big* has the maximal stress in the sentence. In turn, the metrical grid is bracketed into units that represent the *prosodic contours* of the utterance—its division into breath groups over which intonational contours are defined. In figure 1.1, the bracketing indicates a division something like *The LITTLE star's—beside a BIG star*. I have not indicated here the intonation contours themselves; in a tone language such as Mandarin, there would be additional structure indicating the tones associated with each syllable.[2]

So far this is just a structured string of sounds; I've said nothing about the division of the string of sounds into words! This division appears below the segmental structure as another sequence of trees (which for convenience are notated upside down), the *morphophonology*. These trees say that the sentence has five full phonological words: *little*, *star*, *beside*, *big*, and *star*. Attached to some of them are clitics, corresponding to *the*, *'s*,

[2] Influential treatments of intonation contours include Pierrehumbert 1980, Beckman and Pierrehumbert 1986, and Ladd 1996; for tone languages, see Yip 1995. More generally for phonology, see Goldsmith 1995.

and *a*. Notice that the syllabic structure and the morphophonology don't match up exactly. In particular, the clitic *'s* forms part of a syllabic coda with the last consonant of *star*.

All this structure so far is phonology. It says nothing about parts of speech such as nouns and verbs. These categories appear in *syntactic structure*, the next major part of figure 1.1. I have notated this as a tree structure of more or less the familiar sort. There is one important difference: the words *the*, *little*, *star*, and so forth are not notated in the syntactic tree in the conventional fashion. My reason for doing it this way is developed in detail in chapter 2. For now, the basic point is to segregate the different kinds of linguistic features into their proper structures. In particular, the fact that the word is pronounced *star* is a fact of phonology, not of syntax. All the *syntax* knows is that it is a noun, indistinguishable from every other singular count noun in English (in languages such as French, Russian, and Hebrew, grammatical gender would also be notated here).

However, the overall structure must of course indicate that the phonological piece *star* corresponds to a noun in syntactic structure; this is notated with the letter subscripts in figure 1.1. For instance, the subscript *e* connects the word *star* in morphophonology with the first noun in the syntax. Look also at the clitic *z* next to *star*, with the subscript *f*, which is linked to the inflected verb of the sentence. This little *z* is thus the phonological encoding of the verb *be* in present tense, inflected for third person singular—in other words, the contracted form of *is*.

We've still said nothing about what the sentence means. This is the role of the two structures at the bottom of figure 1.1. The *semantic/conceptual structure* is an algebraic encoding of the propositional organization of the sentence, in function-argument form—a predicate calculus sort of structure. It's over this structure that principles of inference, reference, and truth-conditions can be defined formally. In this particular example, there is a Situation in the present, which consists of a State of a Thing being in a Place. The Thing is of the category STAR, it has the property of being LITTLE, and it is definite (i.e. the speaker takes it to be independently identifiable by the hearer of the utterance). The Place (where the little star is) is a region of space that is determined by a spatial relation, BESIDE, in relation to a reference object. In turn, the reference object is also of the category STAR, it has the property of being BIG, and it is indefinite—that is, it is an entity new to the discourse. (For a little more detail, see section 6.1.)

These pieces of the semantic structure are coindexed with the syntactic structure (and therefore indirectly with the phonology) by number subscripts. For instance, the syntactic subject of the sentence (the first NP) has the index *2*, which corresponds with the first Thing constituent of the semantic structure (i.e. the meaning of the phrase *the little star*). Now notice one particular curious correspondence: the semantic feature PRES (present time) has subscript *7*, so it corresponds to present tense in syntax, a feature of the verb's inflection. But this feature of the verb doesn't correspond directly to anything in phonology. Rather, it is swallowed up as part of the inflected verb, which in turn surfaces as the clitic *'s* in phonology—not even a syllabic coda on its own. Thus the outermost functional element in meaning, the one that provides the whole framework for the meaning, surfaces as only a tiny part of the tiniest part of the phonology. This sort of mismatch turns out not to be so unusual in language.

The semantic/conceptual structure in turn maps in some ill-understood way into a spatial or visual encoding of the scene that the sentence describes, so that the sentence can be used to describe a visual scene. I have notated this crudely as the *spatial structure* in figure 1.1 (one could think of this as the "mental model" of the sentence in Johnson-Laird's (1983) sense, or alternatively as a visual percept or visual image). Here the subscripts connect the parts of the visual figure to their corresponding elements in semantic/conceptual structure. The dashed oval in spatial structure corresponds to the spatial region expressed as *beside the star*—something that is not present in visual phenomenology but *is* present in visual understanding. (Of course, in a sentence expressing an abstract proposition, there will be no corresponding spatial structure.)

This completes our tour of the structure of this ridiculously simple sentence. For more complex sentences like those we use constantly, there will be much more of the same. I want to emphasize that all this structure represents a pretty fair consensus among linguists, based on research on thousands of linguistic phenomena in hundreds of languages of the world. This research includes not only speakers' judgments of grammaticality but also analysis of texts, historical change in languages, experimental psycholinguistic research on online processing in perception and production, the acquisition of language by children and adults, the loss of language by aphasics, and so on. I stress the motivation for the analysis because people outside of linguistics sometimes think that linguists just make all this up. Nothing could be farther from the truth: it's the outcome of rigorous empirical research.

1.4 Relevance to Neuroscience

But what does this structure mean—or what *should* it mean—to a neuroscientist? Of course, there are no symbols like NP and σ running around in our heads. Rather, I think the proper way to understand figure 1.1 is as a claim that there are *functional equivalents* of every element of this structure in our heads. Because this sentence is being produced or understood online, the functional equivalents of these structures must be present in both the speaker's and the hearer's working memory. A sentence is not just a string of words, each of them being a node in a semantic network or some such. It is a set of three or more correlated structures: phonology, syntax, semantics, and (sometimes) spatial structure, each of which has its own particular dimensions of variation, its own repertoire of basic elements, and its own principles of combination. In producing a sentence, one must map from a semantic structure (the meaning one wishes to express), through syntax, to phonology, which leads to the formation of instructions to the vocal tract. In hearing and understanding a sentence, one must convert an acoustic signal into phonology, which in turn can be mapped to syntactic and semantic structures in working memory.[3] Language processing cannot go directly from acoustics to meaning or from meaning to motor control, because the correspondence is determined by the principles of the language: think again of how the meaning 'present time' is related to phonological expression only as a part of the meaning of the little sound *z*. And in the course of producing or understanding the sentence, the speaker and hearer need all these structures to be available simultaneously in working memory, as is clear from the fact that they know which words correspond to which parts of the meaning.

Naturally, both neuroscientists and linguists would love to know how these structures are instantiated in neural tissue and neural activity. But this is not a question that can be answered at present. In particular, even if we know *where* a structure is localized in the brain—the sort of information that neural imaging can provide—we do not know *how* the brain instantiates the structure. I think it is worth emphasizing our extreme ignorance here. We don't have the slightest idea how even the most elementary units of linguistic structure such as speech sounds can be instantiated

[3] This is an oversimplification, of course. It is not as though one hears a whole sentence, then parses it all syntactically, then decides what it means. Rather, processing is incremental and involves feedback. See section 1.5.2.

neurally: how speech sounds are stored and how they are processed. Some neuroscientists say we are beyond this stage of inquiry, that we don't need to talk about "symbols in the head" anymore. I firmly disagree. We *know* that language is organized into speech sounds and that speech sounds are only the first step in analyzing linguistic structure. As far as I know, there exist absolutely no attempts to account for even this trivial degree of linguistic complexity in neural terms, and speech sounds only scratch the surface. In my opinion, it is the height of scientific irresponsibility to totally dismiss linguistic theory, claiming that some toy system (say a computational neural network) will eventually scale up to the full complexity of language.[4] A linguist who made comparably ignorant claims about the brain would be a laughingstock. End of sermon.

The structure in figure 1.1 tells us still more about how the brain has to be functionally organized. First, consider the subscripting that connects the structures to each other. This presents an especially complex example of the familiar *binding problem* in neuroscience, a term usually applied to the problem of connecting different aspects of visual representations such as motion, color, and shape, which are (I gather) processed in different brain areas (Treisman 1988). Figure 1.1 shows how to connect different aspects of linguistic representations: sound, grammatical structure, and meaning. What is striking here is that this trivial little sentence requires a staggering amount of binding: each of the 23 subscripts represents a different pair of pieces that has to be connected—simultaneously. Any moderately complex sentence, such as the one you are now reading, requires vastly more binding. This presents a challenge. I gather that the most popular hypothesis for binding is that bound constituents fire in temporal synchrony with each other and out of synchrony with other elements (e.g. Gray et al. 1989; Crick and Koch 1990; Singer et al. 1997). But can this account cope with the binding in figure 1.1? Could there be enough temporal bandwidth in neural firing to discriminate 23 separate bindings at once? This problem is not particular to language, of course; similar problems of massive binding will arise for the integration of any visual scene of medium complexity.

A different challenge for binding arises from the fact that there are two occurrences of the word *star* in the sentence in figure 1.1. Presumably, long-term memory contains one copy of this word. Yet the word *star* must be bound (or copied) to two separate locations in working memory,

[4] This is approximately the gist of Elman et al. 1996 and Deacon 1997, for instance.

and both copies must be simultaneously present and active in working memory in order for the sentence to be produced, understood, and connected with the visual percepts. But the two copies had better not be bound together; if they were, we would understand there to be a single star that is both little and big! Again, this is not a particularly linguistic problem; it arises any time there are two tokens of the same category in a visual configuration, for instance two identical coins on a table. But the linguistic case points up the essential nature of the problem. The problem, of course, is the necessity for mental representation to be able to discriminate types, stored in long-term memory, from tokens, instantiated in working memory. This is an issue discussed at length by Marcus (2001) in his critique of the most popular variety of connectionist learning.

These issues are treated in more detail in Jackendoff 2002a, especially chapter 3. The message to take from the present discussion is that an investigation of mental structures provides important boundary conditions on the theory of brain function. A similar point was made by Marr (1982) in connection with vision. In both language and vision, if we want to figure out how the brain works, it behooves us to try to understand what functions the mind has to compute. A proposed theory of neural behavior is incomplete if it does not offer genuine solutions to the problems of combinatoriality, structural hierarchy, and binding among structures.

It is not that these problems are particular to language. It is just that linguistic theory focuses on these problems and builds on them in a way that theories of other "vertical" faculties of mind usually have not. Part of the message of this book is that these properties recur in other faculties, should we care to look for them. Sixty years ago, nearly everyone thought that language was perfectly transparent and hardly complex at all (and many nonlinguists, even some in psychology and neuroscience, still think so). Since then we have learned that not only is language far more complex than we ever would have dreamed, but so is every other aspect of the mind/brain that has been investigated.

To sum up: Pretty much all cognitive neuroscientists agree in rejecting dualism; ultimately the mind must run in the brain, and there are no mental properties that are causally independent of brain events. However, to insist that neural accounts have absolute priority, that they somehow have a greater reality or are "more scientific" than functional accounts, to me has a chilling effect on inquiry. It seems to me that in the practice of research, the relationship between neural and functional accounts ought to be a two-way street: what we know about each dimension of the problem ought to enrich our study of the other.

1.5 An Overall Vision of Mental Architecture

1.5.1 Levels of Structure and Interfaces

Extrapolating from linguistic theory, a vision of an overall "vertical" architecture of mind emerges. The division of the mind into "faculties" and their "subfaculties" is instantiated by a collection of discrete levels of structure, of which phonology, syntax, conceptual structure, and spatial structure are "subfaculties" involved in the language faculty. Each of these levels has its own characteristic basic elements (e.g. distinctive features and syllabic units in phonology) and its own characteristic combinatorial principles (e.g., in phonology, collection of features into segments and concatenation of segments into syllables). In addition, mental structure is governed by *interface* principles that connect particular pairs of levels (or perhaps larger *n*-tuples of levels). Such principles connecting levels L_1 and L_2 establish which parts of L_1 correspond to which parts of L_2; the corresponding parts are bound, as indicated by the subscripts in figure 1.1. A leading question of cognitive science therefore ought to be this:

- What are the levels of mental structure, and what are the interfaces among them?

Notice next that the levels of conceptual structure and spatial structure do not belong to the language faculty per se: they play a role in many different faculties, including vision and action. In contrast, phonology and syntax *are* specific to the language faculty and therefore might be considered (part of) the "narrow language faculty" in the sense of Hauser, Chomsky, and Fitch 2002.[5] The interfaces through which the "narrow language faculty" communicates with conceptual structure and spatial structure are qualitatively not unlike the interface between phonology and syntax: in each case, the interface establishes a correlation between parts of structures.

More broadly, the question arises of how one can talk about what one sees: how the visual faculty communicates with the language faculty. The answer is that the visual faculty comprises a collection of levels connected by interfaces, of which the most peripheral are the distinctions made by the retina and primary visual cortex, and among the most central is the

[5] Though Hauser, Chomsky, and Fitch do not consider phonology part of the narrow faculty, as they make clear in Fitch, Hauser, and Chomsky 2005, in response to Pinker and Jackendoff 2005.

level of spatial structure. In turn, spatial structure has interfaces that lead into the language faculty. In other words, multimodal interactions are made possible by interfaces that link levels used by the different faculties.

Looking at all the levels "horizontally," we might notice that many different levels of structure are hierarchical, in that elements of structure are combined to make higher-order elements, which in turn combine with other elements. We might further notice that some levels are recursive, in the sense that a structural element of a particular type can form a constituent of another element of the same type. For example, syntactic structure is recursive, in that an element of the type Noun Phrase can be a constituent of another Noun Phrase, as in *[$_{NP}$ the king of [$_{NP}$ the Cannibal Islands]]*, and this embedding can be repeated, as in *[the tip of [the nose of [the father of [the bride of [the king of [the Cannibal Islands]]]]]]*. On the other hand, syllabic structure, though hierarchical, is not recursive, in that such unrestricted embedding is not possible. So a more general question arises:

- Which levels of structure are hierarchical, and, among those, which are recursive?

Of course, this question cannot be answered in a principled way until we have accounts of numerous levels of structure in different faculties. Hauser, Chomsky, and Fitch (2002) speculate that recursion may be unique to humans and in particular to the language faculty, perhaps even *the* single factor that makes language a human specialization; then they back off and speculate that recursion might be found elsewhere in cognition. Jackendoff and Pinker (2005) confirm this speculation, pointing out that figure 1.3 shows evidence of recursion in visual cognition. This display is perceived as being built recursively out of discrete elements that combine to form larger discrete constituents: pairs of *x*s, clusters of four pairs, squares of four clusters, arrays of four squares, arrays of four arrays, and so on. One could further combine four of these superarrays into a still larger array, and continue the process indefinitely. So, to use Chomsky's term, we have here a domain of "discrete infinity" in visual perception, with hierarchical structure of unlimited depth, its organization in this case governed by classical Gestalt principles. Presumably the principles that organize figure 1.3 play a role in perceiving objects in terms of larger groupings, and in segregating individual objects into parts, parts of parts, and so on. Similar principles of grouping appear in music (Lerdahl and Jackendoff 1983).

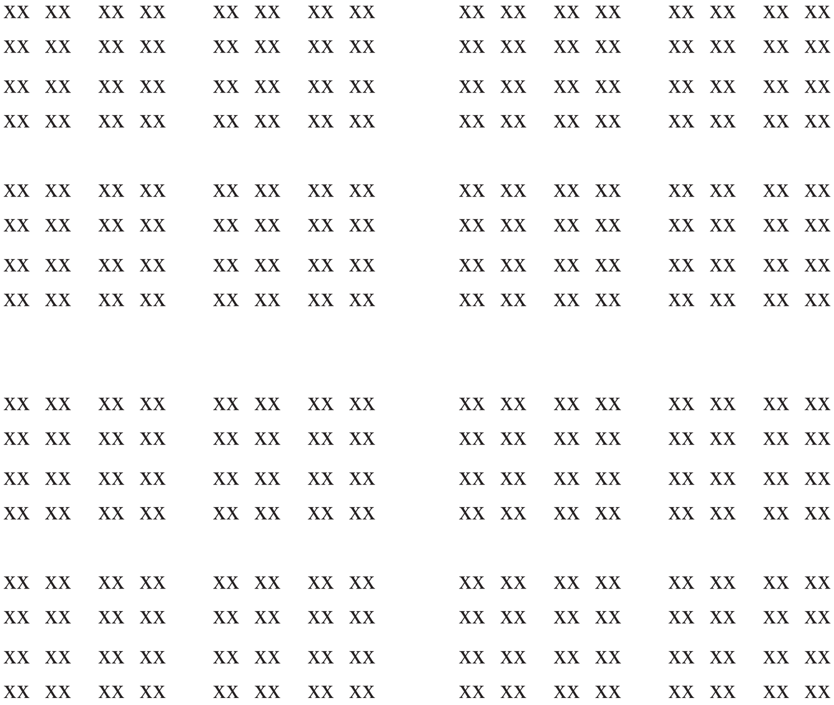


Figure 1.3
Recursion in visual grouping

When the similarities and the differences are examined, it appears that hierarchical phrase structure in language cannot be reduced to the principles governing visual and musical grouping. Two formal properties distinguish recursion in syntax. First, elements and phrases of syntax belong to distinguishable syntactic categories such as N or VP; visual groups do not obligatorily fall into some small set of distinguishable categories (as far as we know). The particular family of categories in syntactic phrases appears to be sui generis to syntax. Second, unlike what we find in visual grouping, one member of each syntactic constituent has a distinguished status as *head*, such that the other members are considered dependent on it. Headed hierarchies are found elsewhere in cognition, for instance in syllabic structure (which, as mentioned, is not recursive in the strong sense), in conceptual structure, in certain aspects of musical structures (Lerdahl and Jackendoff 1983; Jackendoff 1987, 249–251), and, as I will argue in chapter 4, in the structure of complex action.

1.5.2 Processing

A theory of structure alone is not a theory of mental functioning. It must be complemented with a theory of how mental structures are processed over time to produce behavior, knowledge, and experience. In present terms, the basic processing operations are the construction of mental structures at each level and the linking of structures at multiple levels. Consider for example language perception. Environmental input leads to the construction[6] of an auditory structure. The interface that links this to phonology leads to the construction of a candidate phonological structure in working memory. The further interfaces to syntax and thence to conceptual structure lead to construction of an interpretation for the heard utterance, also in working memory.[7] However, this process of construction is not just an autonomous function of working memory. In order to get from a phonological string to a meaning, the processor must call on material stored in long-term memory. In particular, the words of the utterance must be identified in order to assign particular chunks of phonology to chunks of meaning. If the hearer doesn't know the words, the meaning cannot be determined. If language perception is successful, the outcome is a set of structures linked in working memory (where the linkings are notated in figure 1.1 as the matched subscripts). In turn, one or more of these structures may be shipped to long-term memory—if only the conceptual structure, one remembers the gist; if all the structures, one has memorized the sentence.

Pretty much everyone imagines similar processes of construction emerging in visual perception, with structure propagating up from sensory to central levels. The main doubt that might arise is whether visual

[6] Or in this particular case, "transduction," in the sense of Pylyshyn 1984. The process by which sensory stimulation gives rise to a functional organization in the mind cannot be characterized in functional terms—only its output can. In other words, this marks the outer boundary of the applicability of functional description.

[7] I am taking working memory to be an active "workbench" or "blackboard" on which mental structures are constructed and manipulated, rather than just a passive store for rehearsal in the sense of Baddeley 1986, for instance. I gather my sense is not universally accepted. Readers should feel free to substitute their own favorite term for the functional capacity that builds sentence structures (other than "central executive," for if Fodor's (1983) notion of modularity is right about anything, it is that sentence perception is not a central executive function). See Jackendoff 2002a, sec. 7.3.

perception draws on anything akin to the learned store of words; the general presumption is that visual processing employs only more general (and unlearned) principles. However, some researchers (see e.g. Marr 1982; Ullman 1998; Cavanagh, Labianca, and Thornton 2001) have proposed that some higher-order visual processing is mediated in part by learned familiarity with certain kinds of objects and motions. Moreover, at least at the most central levels, much learning is necessary to establish cross-modal connections. For instance, everyone has learned an association between appearance and taste for hundreds if not thousands of kinds of food; we might consider this cross-modal knowledge a kind of "visual-to-taste lexicon" that helps interface between the two sorts of mental structures. Chapter 4 further proposes that there is a large "action lexicon," which encodes learned complex actions for purposes of both production and perception, and which links artifacts (e.g. doors, coffeepots, faucets) with the appropriate actions performed using them.

A further kind of processing has to be mentioned: processes in which a structure on a particular level leads to construction of new structure(s) on the same level. The most prominent case is reasoning, which builds new conceptual structures from old. Such construction is governed by principles of inference (both logical and heuristic), which in this perspective are mappings from conceptual structures into further conceptual structures. But other mental processes might be treated as similar types of "within-level" construction, for instance mental rotation, which manipulates visual structures, and the computation of rhyme, which compares phonological structures.

Some general properties of the process of construction have emerged in research on language processing and also, I believe, in research on vision. First, construction is *incremental*: one does not need to flesh out a whole level before proceeding to the next. Rather, as soon as some structure is present in level L_1 that can be correlated with structure in the next level L_2, the interface linking L_1 and L_2 instantiates the correlated structure in L_2.

Second, construction is *promiscuous*: often the structure at a particular level is underdetermined by the process of construction up to that point in time. In general, the processor does not arbitrarily choose among the possibilities and then go on from there (as in the algorithmically conceived processing theories prevalent in the 1970s). Rather, it constructs *all* reasonable possibilities and runs them in parallel, eventually selecting a single most plausible or most stable structure as more constraints become available, and inhibiting the other structures. For instance, Swinney

(1979) and Tanenhaus, Leiman, and Seidenberg (1979) demonstrate that when a word in a sentence is first heard, all of its possible meanings are activated, whether contextually plausible or not; the otiose meanings are pared away over time as the word is integrated with the syntactic, semantic, and pragmatic context.[8]

Third, in order for such contextual effects to be possible, perception cannot be just bottom up. Rather, there has to be a degree of interaction in both directions. I find it useful to think of the process of construction as achieving a "resonance" among the linked structures, a state of global optimal stability within and among the structures in the complex. Occasionally among the promiscuous structures there are multiple stable states, in which case perception produces an ambiguous result such as the Necker cube in vision and a pun or other ambiguity in language.

Fourth, the processes of propagation through interfaces can in many cases be run in either direction. For instance, language perception is the process of beginning with a phonological structure and propagating structure to conceptual structure; and language production is the opposite, beginning with something to say (a conceptual structure) and propagating structure to phonology. The only part of the process that is unidirectional is the very periphery: one goes from audition to phonology and not the other way about, and one goes from phonology to motor control and not the reverse.

In vision, one is accustomed to thinking only in terms of perception, hence propagation of structure from sensory to central levels. But an instruction to imagine an elephant can provoke visual imagery. In such a case, the construction has to proceed from the interface(s) of the visual faculty with the most central levels of language structure. To the extent that noncentral levels of vision are involved in visual imagery (say if primary visual cortex is shown to be activated), propagation of activation has to be top-down. This means that except at the very periphery, visual processing too can be bidirectional.

1.5.3 Learning

A theory of mental function must be concerned not just with processing but also with learning. There are at least two different cases. An example of the first has already been mentioned: the taking in of information

[8] The notion of promiscuity also shows up in Dennett's (1991) idea of the mind as constructing "multiple drafts," only one of which is selected to be the "narrative" in terms of which one understands and remembers one's current situation.

conveyed by an utterance. More generally, this is what goes under the term "one-time learning"—committing to long-term memory a structure that has been constructed in working memory. The formation of episodic memories would also fit under this rubric, if only we had a theory of the mental structures involved in perceiving and understanding "episodes."

The other type of learning, which might be called "slow(er) learning," involves the consolidation and generalization of material in long-term memory into schemas. For example, one approach to learning "rules of grammar" (e.g. Goldberg 1995, 2006; Culicover 1999; Jackendoff 2002a, secs. 6.9–6.10; Tomasello 2003; Culicover and Jackendoff 2005) conceives of them as abstractions from the structures of actual sentences the language learner has experienced. The formal difference between actual sentence structures and rules of grammar is that rules contain variables to be instantiated; the utterances one has experienced represent various instantiations of these variables. In other words, to learn a rule is to extract commonality among instances and replace the differences among the instances with a variable. In turn, rules of similar form can be further generalized, resulting in a stored schema with more and/or broader variables.

The result is a long-term memory that is more than a list of memories: it is structured in terms of "inheritance hierarchies," in which stored instances are at the bottom of the hierarchy and the most general schemas are at the top. This corresponds to a fairly broadly accepted sense of "semantic memory." However, in the "item-based" approach to language acquisition, inheritance hierarchies can be applied not only to semantic schemas such as *poodle → dog → animal → living thing*, but also to purely formal syntactic and phonological structures.

I take it that this type of "slow learning" is a process that takes place within long-term memory—as it were, a constant resifting of experience behind the scenes. But not much hangs on this.

1.6 A Caution, and What Modularity Means

It is common in cognitive neuroscience circles to speak of "information being broadcast through the brain." Here are some typical statements of this sort:

> ...the contents of awareness are to be understood as those information contents that are accessible to central systems, and brought to bear in a widespread way in the control of behavior. (Chalmers 1997, 22)

> . . . conscious contents become ‘globally available’ to many unconscious systems. The reader’s consciousness of *this phrase*, for example, makes *this phrase* available to interpretive systems that analyze its syntax and meaning, its emotional and motivational import, and its implications for thought and action. (Baars 1997, 241)

> . . . it seems reasonable to hypothesize that awareness of a particular element of perceptual information must entail . . . access to that information by most of the rest of the mind/brain. (Kanwisher 2001, 105)

> . . . dynamic mobilization makes [information available within a modular process] directly available in its original format to all other workspace processes. (Dehaene and Naccache 2001, 15)

On the view of mental structure and function being advocated here, this notion cannot be sustained. A phonological structure, for example, is intelligible “in its original format” only to the part of the mind/brain that processes phonological structure. If that part of the mind “broadcast” its contents to, say, a visual processor, it would be less than useless. And the same is true for any level of structure.

There is however a more restricted sense in which information is “broadcast.” To the extent that a level of structure has interfaces to other levels, the interfaces can propagate activation to the related levels—but in the levels’ own proprietary formats. So, for instance, phonological structure has an interface with syntax, so the presence of a phonological structure in working memory leads fairly automatically to the construction of a correlated syntactic structure. In turn, the syntactic structure interfaces with conceptual structure, so a linked triple of structures emerges over time. If the conceptual structure turns out to be an instruction to form a mental image, understanding the sentence leads to the propagation of structure into the visual system as well—in visual, not phonological format. But getting to the visual format requires passing through all the intervening interfaces.[9]

This is the sense in which I want to understand the notion of modularity. Each level of structure has its own proprietary format, incompatible with all the others. Thus it is “domain-specific” in the sense of Fodor 1983. Its interfaces with other levels of structure are what prevent it from

[9] An exception: There is a specialized interface between phonology and vision that is responsible for *reading*. Thus a phonological structure *[elephant]* in working memory might lead to the construction of a visual image of the corresponding written form. The image of an elephant, however, would have to come by the normal route.

being functionally isolated in the mind. This leads to a relativized version of Fodor's notion of "informational encapsulation": a level L_1 is encapsulated from another level L_2 to the degree that distinctions in L_1 do not have direct correlates in L_2. For instance, phonological structure is less encapsulated from syntax than it is from spatial structure, in that phonological linear order and constituency correspond fairly closely with syntactic linear order and consituency, whereas the relation between phonology and spatial structure is far less direct, mediated by several intervening interfaces. (See Jackendoff 2002a, sec. 7.5, for more detailed comparison of this "structure-based" modularity with Fodorian modularity.)

I emphasize that all of this discussion of mental structure, outside of the language faculty, is strictly programmatic. Some of the later chapters of this book are in part an attempt to demonstrate the utility of this approach in studying other cognitive phenomena.

Chapter 2

Reintegrating Generative Grammar

2.1 Introduction

In the middle 1960s, when I began my graduate study in linguistics at MIT, generative grammar was *the* hot new topic.[1] Everyone was reading about it, from philosophers to psychologists to anthropologists to educators to literary theorists to musicians. It was an amazing time to be in the field, comparable, I imagine, to the early days of nuclear physics, jazz, or impressionist painting. Our padrone, Morris Halle, was in his early forties, Noam Chomsky was in his late thirties, and nearly all the rest of us were in our twenties. We were making up this new wonderful thing by the seat of our pants; we knew all the players personally; there was a marvelous sense of adventure and spirit of collaborative play. And every part of the field was wide open: there was virtually no literature, and most of what there was was pounded out on aged typewriters, mimeographed, collated by hand, and circulated in samizdat (photocopying was still expensive, and PCs and the Web were decades in the future). As my late friend Adrian Akmajian once put it, every time you opened your mouth you made history.

By the late 1970s, though, the bloom was off the rose, even if most linguists didn't realize it; and by the 1990s, linguistics was arguably far on the periphery of the action in cognitive science. To some extent, such a decline in fortune was simply a matter of the inevitable maturing of a field, changes in fashion, and the arrival of new methodologies such as connectionism and brain imaging. However, I believe there are deeper reasons for linguistics' loss of prestige, some historical and some scientific.

[1] This chapter (a revised version of Jackendoff 2003) is based on material developed at much greater length in Jackendoff 2002a and Culicover and Jackendoff 2005. A version of this chapter was presented as my Presidential Address to the Linguistic Society of America in 2003.

The basic questions I want to take up in this chapter, therefore, are these:

- What was *right* about generative grammar in the 1960s, such that it held out such promise?
- What was *wrong* about it, such that it didn't fulfill its promise?
- How can we fix it, so as to restore its value to the other cognitive sciences?

The goal is to integrate linguistics with the other cognitive sciences, not to eliminate the insights achieved by any of them. To understand language and the brain, we need all the tools we can get. But everyone will have to give a little in order for the pieces to fit together properly.

This chapter argues that the overall program of generative grammar was correct, as was the way this program was intended to fit in with psychology and biology. However, several basic technical mistakes at the heart of the formal implementation led to the theory's being unable to make the proper connections both within linguistics and with neighboring fields. This chapter proposes an alternative implementation of the program of generative grammar, the *parallel architecture*, which offers far richer opportunities for integration of the field. In order to understand the motivation for the parallel architecture, it is necessary to look at some history.

2.2 Three Founding Themes of Generative Grammar

The remarkable first chapter of Noam Chomsky's *Aspects of the Theory of Syntax* (1965) set the agenda for everything that has happened in generative linguistics since. Three theoretical pillars support the enterprise: *mentalism*, *combinatoriality*, and *acquisition*.

2.2.1 Mentalism

Before *Aspects*, the predominant view among linguists—if it was even discussed—was that language is something that exists either as an abstraction, or in texts, or in some sense "in the community" (the latter being the influential view held by Saussure (1915), for example). Chomsky urged that the appropriate object of study is the linguistic system in the mind/brain of the individual speaker. According to this stance, a community has a common language by virtue of all speakers in the community having essentially the same linguistic system in their minds/brains.

Note that "essentially the same system" is a matter of perspective. When we are talking about "English speakers" as a whole, we can treat them all as essentially the same. But if we are talking about dialect differences, dialect contact, or language change, we can just as easily switch to treating different speakers as having (slightly) different linguistic systems in their heads. And of course when we are talking about language acquisition, we take it for granted that the young child has a different system than the adult.

The term most often used for this linguistic system is "knowledge," perhaps an unfortunate choice. However, within the theoretical discourse of the time, the alternative was thinking of language as an ability, a "knowing how" in the sense of Ryle 1949, which carried overtones of behaviorism and stimulus-response learning, a sense from which Chomsky with good reason wished to distance himself. It must be stressed, though, that whatever term is used, most of the linguistic system in a speaker's mind/brain is deeply unconscious, unavailable to introspection, in the same way that our processing of visual signals is deeply unconscious. Thus language is a kind of mind/brain property hard to associate with the term "knowledge," which commonly implies accessibility to introspection. We might compromise with tradition by using the term *f-knowledge* ('functional knowledge') to describe whatever is in speakers' heads that enables them to speak and understand their native language(s). Such a term distances linguistic competence from any hints of intentionality in the sense of chapter 1 and makes us concentrate on structure (which is what linguists do anyway).

There still are linguists, especially those edging off toward semiotics and hermeneutics, who reject the mentalist stance and assert that the only sensible way to study language is in terms of communication between individuals (a random example of work taking this view is Dufva and Lähteenmäki 1996). But on the whole, the mentalistic outlook of generative grammar has continued to be hugely influential throughout linguistics and cognitive neuroscience.

More controversial has been an important distinction made in *Aspects* between the study of *competence*—a speaker's f-knowledge of language—and *performance*, the actual processes (viewed computationally or neurally) taking place in the mind/brain that put this f-knowledge to use in speaking and understanding sentences. I think the original impulse behind the distinction was methodological convenience. A competence theory permits linguists to do what they have always done, namely study phenomena like Serbo-Croatian case marking and Turkish vowel harmony, without worrying too much about how the brain actually processes them. Unfortunately, in response to criticism from many different quarters (especially in response to the collapse of the derivational theory of complexity, as detailed in Fodor, Bever, and Garrett 1974, for example), linguists tended to harden the distinction into a firewall: competence theories came to be considered immune to evidence from performance. And so began a gulf between linguistics and the rest of cognitive science that has persisted until today.

We need not abandon the competence/performance distinction, but we should return it to its original status as a methodological rather than ideological distinction. Although the innovations proposed here are largely in the realm of competence theory, one of their important consequences is a far closer connection to theories of processing, as well as the possibility of a two-way dialogue between competence and performance theories. We return to this issue in section 2.10.2.

2.2.2 Combinatoriality

The earliest published work in generative grammar, Chomsky's *Syntactic Structures* (1957), began with the observation that a language contains an indefinitely large number of sentences. Therefore, in addition to the finite list of words, a characterization of a language must contain a set of *rules* (or a *grammar*) that collectively describe or "generate" the sentences of the language. *Syntactic Structures* showed that the rules of natural language cannot be characterized in terms of a finite-state Markov process, nor in terms of a context-free phrase structure grammar. Chomsky proposed that the appropriate form for the rules of a natural language is a context-free phrase structure grammar supplemented by transformational rules. Not all subsequent traditions of generative grammar (e.g. Head-Driven Phrase Structure Grammar (Pollard and Sag 1994) and Lexical-Functional Grammar (Bresnan 1982, 2001)) have maintained the device of transformational rules; but they all contain machinery designed to overcome the shortcomings of context-free grammars pointed out in 1957.[2]

Carried over into the mentalistic framework of 1965, the consequence of combinatoriality is that speakers of a language must have rules of lan-

[2] To some extent, Chomsky's point has been lost on the larger cognitive neuroscience community. For instance, Elman's (1990) widely cited recurrent network parser is a variant of a finite-state Markov device and is therefore subject to some of the same objections raised by Chomsky in 1957. See Marcus 2001 and Pinker 1999 for extensive discussion.

guage (or mental grammars) in their heads as part of their f-knowledge. Again, a certain amount of controversy has arisen from the term "rules." Rules of grammar in the sense of generative grammar are not like any of the sorts of rules or laws in ordinary life: rules of etiquette, rules of chess, traffic laws, or laws of physics. They are unconscious principles that play a role in the production and understanding of sentences. Again, to ward off improper analogies, we might use the term *f-rules* for whatever the combinatorial principles in the head may be. Generative linguistics leaves open how directly the f-rules are involved in processing, but, as suggested above, the unfortunate tendency among linguists has been not to care. The approach advocated here, though, makes it possible to regard the rules as playing a direct role in processing (see again section 2.10.2).

An important reason for the spectacular reception of early generative grammar was that it went beyond merely claiming that language needs rules: it offered rigorous formal techniques for characterizing the rules, based on approaches to the foundations of mathematics and computability developed earlier in the century. The technology suddenly made it possible to say lots of interesting things about language and ask lots of interesting questions. For the first time ever, it was possible to provide detailed descriptions of the syntax of natural languages (by 1965, generative grammarians had studied not only English but German, French, Turkish, Mohawk, Hidatsa, and Japanese as well). In addition, generative phonology took off rapidly, adapting elements of Prague School phonology of the 1930s to the new techniques. With Chomsky and Halle's (1968) *Sound Pattern of English* as its flagship, generative phonology quickly supplanted the phonological theory of the American structuralist tradition.

2.2.3 Acquisition

Mentalism and combinatoriality together lead to the crucial question: how do children get the f-rules into their heads? Given that the f-rules are unconscious, parents and peers cannot verbalize them; and even if they could, children would not understand, since they don't know language yet. The best the environment can do for a language learner is to provide examples of the language in a context. From there on, it is up to the language learner to construct the principles on his or her own—unconsciously of course.

Chomsky asked the prescient question: what does the child have to "(f-)know in advance" in order to accomplish this feat? He phrased the problem in terms of the "poverty of the stimulus": many different generalizations are consistent with the data presented to the child, but the child somehow comes up with the "right" one—the one that puts him or her in tune with the generalizations of the language community. I like to put the problem a bit more starkly. The whole community of linguists, working together for decades with all sorts of crosslinguistic and psycholinguistic data unavailable to children, has still been unable to come up with a complete characterization of the grammar of a single natural language. Yet every normally developing child does it by the age of 10 or so. Children don't have to make the choices we linguists do: for instance, they don't have to decide whether the "right" choice of grammar is in the style of transformational grammar, the Minimalist Program, Optimality Theory, Role and Reference Grammar, Tree-Adjoining Grammar, Cognitive Grammar, connectionist networks, or some as yet unarticulated alternative. They already f-know it in advance.

One of the goals of linguistic theory, then, is to solve this "paradox of language acquisition" by discovering what aspects of linguistic f-knowledge are *not* learned, but rather *form the basis* for the child's learning. The standard term for the unlearned component is *Universal Grammar* or *UG*, a term that, like "knowledge," perhaps carries too much unwanted baggage. In particular, UG should not be confused with universals of language: UG is rather what shapes the acquisition of language. I prefer to think of it as a toolkit for constructing language, out of which the child (or better, the child's mind/brain) f-selects tools appropriate to the job at hand. If the language in the environment happens to have a case system (like German), UG will help shape the child's acquisition of case; if it has a tone system (like Mandarin), UG will help shape the child's acquisition of tone. But if the language in the environment happens to be English, which lacks case and tone, these parts of UG will simply be silent.

What then is the source of language universals? Some of them will indeed be determined by UG, for instance the overall "architecture" of the grammatical system: the parts of the mental grammar and the relations among them (of which much more below). Other universals, especially what are often called "statistical" or "implicational" universals, may be the result of biases imposed by UG. For instance, UG may say that if a language has a case system, the simplest such systems are thus-and-so. These will be widespread systems crosslinguistically; they will be acquired earlier by children; and more complex systems may tend to change toward them over historical time. Other universals may be a consequence of the functional properties of any relatively efficient communication system: for instance, the most frequently used signals tend to be short. UG doesn't have to say anything about these universals at all; they will come about through the dynamics of language use in the community (a process that of course is not very well understood).

If UG is not learned, how does the child acquire it? The only alternative is through the structure of the brain, which is determined through a combination of genetic inheritance and the biological processes resulting from expression of the genes, the latter in turn determined by some combination of inherent structure and environmental input. Here contemporary science is pretty much at an impasse. We know little about how genes determine brain structure and (as emphasized in chapter 1) nothing about how the details of brain structure determine anything about language structure, even aspects of language as simple as speech sounds. Filling out this part of the picture is a long-term challenge for cognitive neuroscience. Some (see e.g. Elman et al. 1996; Deacon 1997) have rejected the hypothesis of UG, on the grounds that we don't know how genes could code for language acquisition. But such a conclusion is premature. After all, we don't know how genes code for birdsong or sexual behavior or sneezing either, but we don't deny that there is a genetic basis behind these.

Next the question arises of how much of UG is a human cognitive specialization for language and how much is a consequence of more general capacities. The question has often been oversimplified to a binary decision between language being entirely special and language being entirely general, with a strong bias inside generative linguistics toward the former and outside generative linguistics toward the latter. The truth undoubtedly lies somewhere in between. Hauser, Chomsky, and Fitch (2002) propose thinking in terms of a "broad language faculty," the entire collection of mental capacities involved in learning and processing language, and a "narrow language faculty," the aspect of language that is special to language. They offer one hypothesis about how the work is divided up (see section 1.5.1); Pinker and Jackendoff (2005) develop another.

Many people (including myself) would find it satisfying if a substantial part of language acquisition were a consequence of general human cognitive factors. But the possibility that some specializations overlay the general factors must not be discounted. My view is that we cannot really determine what is general and what is special until we have comparable theories of other cognitive capacities, including other *learned* cognitive capacities. To claim that language is parasitic on, say, motor control, perhaps because both have hierarchical and temporal structure (this seems to be the essence of Corballis's (1991) position)—but without stating a theory of the f-knowledge involved in motor control—is to coarsen the fabric of linguistic theory to the point of unrecognizability. One comparable theory is the music theory proposed by Lerdahl and Jackendoff (1983; also Jackendoff and Lerdahl 2006), which displays some striking parallels with language as well as some striking differences. Chapter 4 of this book ventures a start on another faculty, the capacity for complex action, with some interesting consequences.

Of course, if UG—the ability to learn language—is in part a human cognitive specialization, it must be determined by some specifically human genes, which in turn must have come into existence sometime since the hominid line separated from the other great apes. One would therefore like to be able to tell some reasonable story about how the language faculty could be shaped by natural selection or other evolutionary processes. We return to this issue in section 2.10.4.

This approach to the acquisition of language has given rise to a flourishing tradition of developmental research (references far too numerous to mention) and a small but persistent tradition in learnability theory (e.g. Wexler and Culicover 1980, Baker and McCarthy 1981). And certainly, even if the jury is still out on the degree to which language acquisition is a cognitive specialization, all manner of phenomena have been investigated that bear on the issue. For instance:

- The sensitive period for language learning and the consequences for first and second language acquisition at a later age (Lenneberg 1967; Curtiss 1977; Flynn and O'Neil 1988; Newport 1990; Klein and Perdue 1997)
- The limited ability of apes to acquire even rudimentary versions of human language, even with extensive training (Premack 1976; Seidenberg and Petitto 1978; Terrace 1979; Savage-Rumbaugh, Shanker, and Taylor 1998)
- The characteristic brain localization of language functions, resulting in characteristic aphasias (Zurif 1990)
- The grammatical parallels between spoken and signed languages and the parallels in acquisition and aphasia (Klima and Bellugi 1979; Bellugi, Poizner, and Klima 1989; Fischer and Siple 1990)
- The existence of characteristic language deficits associated with various genetic conditions (Bellugi, Wang, and Jernigan 1994; Clahsen and Almazan 1998; Gopnik 1999)
- The creation of creole languages by communities of pidgin-speaking children (Bickerton 1981; DeGraff 1999)

- Most strikingly, the creation of signed languages de novo by a newly assembled community of deaf children, both in Nicaragua (Kegl, Senghas, and Coppola 1999) and in a community of Israeli Bedouins (Sandler et al. 2005)

My impression is that, while there are questions about all of these phenomena, en masse they offer an overwhelming case for some degree of genetic specialization for language learning in humans.

These three foundational issues of generative grammar—mentalism, combinatoriality, and acquisition—have stood the test of time; if anything, they have become even more important over the years within the context of cognitive science. It is these three issues that connect linguistics intimately with psychology, brain science, and genetics. Much of the promise of generative linguistics arose from this new and exciting potential for scientific unification.

2.3 The Broken Promise: Deep Structure Would Be the Key to the Mind

A fourth major point of *Aspects*, and the one that attracted most attention from the wider public, concerned the notion of Deep Structure. A basic claim of the 1965 version of generative grammar was that in addition to the surface form of sentences (the form we hear), there is another level of syntactic structure, called Deep Structure, which expresses underlying syntactic regularities of sentences. For instance, a passive sentence like (1a) was claimed to have a Deep Structure in which the noun phrases are in the order of the corresponding active (1b).

(1) a. The bear was chased by the lion.
    b. The lion chased the bear.

Similarly, a question such as (2a) was claimed to have a Deep Structure closely resembling that of the corresponding declarative (2b).

(2) a. Which martini did Harry drink?
    b. Harry drank that martini.

In the years preceding *Aspects*, the question arose of how syntactic structure is connected to meaning. Following a hypothesis first proposed by Katz and Postal (1964), *Aspects* made the striking claim that the relevant level of syntax for determining meaning is Deep Structure.

In its weakest version, this claim was only that regularities of meaning are most directly encoded in Deep Structure, and this can be seen in (1) and (2). However, the claim was sometimes taken to imply much more:

that Deep Structure *is* meaning, an interpretation that Chomsky did not at first discourage.[3] And this was the part of generative linguistics that got everyone really excited—for if the techniques of transformational grammar could lead us to meaning, we would be in a position to uncover the nature of human thought. Moreover, if Deep Structure were innate—being dictated by UG—then linguistic theory would give us unparalleled access to the essence of human nature. No wonder everyone wanted to learn linguistics.

What happened next was that a group of generative linguists, led by George Lakoff, John Robert Ross, James McCawley, and Paul Postal, pushed very hard on the idea that Deep Structure should directly encode meaning. The outcome, the theory of Generative Semantics (e.g. McCawley 1968; Postal 1970; Lakoff 1971), increased the "abstractness" and complexity of Deep Structure, to the point that the example *Floyd broke the glass* was famously posited to have eight underlying clauses, each corresponding to some feature of the semantics (for the curious, the structure is laid out in Culicover and Jackendoff 2005, sec. 3.3). All the people who admired *Aspects* for what it said about meaning just adored Generative Semantics, and the newer theory swept the country.

But Chomsky himself reacted negatively, and with the aid of his then-current students (full disclosure: present author included), he argued vigorously against Generative Semantics. When the dust of the ensuing "linguistics wars" cleared around 1973 (Newmeyer 1980; Harris 1993; Huck and Goldsmith 1995), Chomsky had won (as usual)—but with a twist: he no longer claimed that Deep Structure was the sole level that determines meaning (Chomsky 1972). Then, with the battle over, he turned his attention, not to meaning, but to relatively technical constraints on movement transformations (e.g. Chomsky 1973, 1977).

The reaction in the larger community was shock: for one thing, at the fact that the linguists had behaved so badly; but more substantively, at the sense that there had been a "bait and switch." Chomsky had promised Meaning with a capital M and then had withdrawn the offer. Many researchers, both inside and outside linguistics, turned away from generative grammar with distaste, rejecting not only Deep Structure but also mentalism, innateness, and sometimes even combinatoriality. And when, later in the 1970s, Chomsky started talking about meaning again, in

[3] For example: "The deep structure that expresses the meaning is common to all languages, so it is claimed [by the Port-Royal grammarians—who of course did not use the term 'deep structure'], being a simple reflection of the forms of thought" (Chomsky 1966, 35).

terms of a syntactic level of Logical Form (e.g. Chomsky 1981), it was too late; the damage had been done. From this point on, the increasingly abstract technical apparatus of generative grammar was of no interest to more than a tiny minority of cognitive scientists, much less the general public. (See Culicover and Jackendoff 2005, chaps. 2 and 3, for a history of these developments and why they drove linguistics away from the rest of cognitive science.)

Meanwhile, various non-Chomskyan traditions of generative grammar were being developed, most notably Relational Grammar (Perlmutter 1983), Head-Driven Phrase Structure Grammar (Pollard and Sag 1987, 1994; Ginzburg and Sag 2000), Lexical-Functional Grammar (Bresnan 1982, 2001), Formal Semantics (Partee 1976; Heim and Kratzer 1998), Optimality Theory (Prince and Smolensky 1993), Construction Grammar (Fillmore, Kay, and O'Connor 1988; Zwicky 1994; Goldberg 1995, 2006), Role and Reference Grammar (Van Valin and LaPolla 1997), Tree-Adjoining Grammar (Frank and Kroch 1995), and Cognitive Grammar (Lakoff 1987; Langacker 1987; Talmy 2000). On the whole, these approaches to linguistics (with the possible exception of Cognitive Grammar) have made even less contact with philosophy, psychology, and neuroscience than the recent Chomskyan tradition. My impression is that many linguists have simply returned to the traditional concerns of the field: describing languages, using whatever theoretical framework they happen to be trained in, and with as little overriding theoretical and cognitive baggage as possible. While this is perfectly fine—particularly since issues of innateness don't play too big a role when you're trying to record an endangered language before all its speakers die—the sense of excitement and danger that comes from participating in the integration of fields has become attenuated.

2.4 A Scientific Mistake: Syntactocentrism

So much for pure intellectual history. We now turn to what I think was an important mistake at the core of generative grammar, one that in retrospect lies behind much of the alienation of linguistic theory from the cognitive sciences. Chomsky did demonstrate that language requires a generative system that makes possible an unlimited variety of sentences. However, he explicitly assumed, without argument (1965, 16, 17, 75, 198), that generativity is localized entirely in the syntactic component of the grammar—the construction of phrases from words—and that phonology (the organization of speech sounds) and semantics (the organization of meaning) are purely "interpretive"; that is, he assumed that their combinatorial properties are derived strictly from the combinatoriality of syntax.

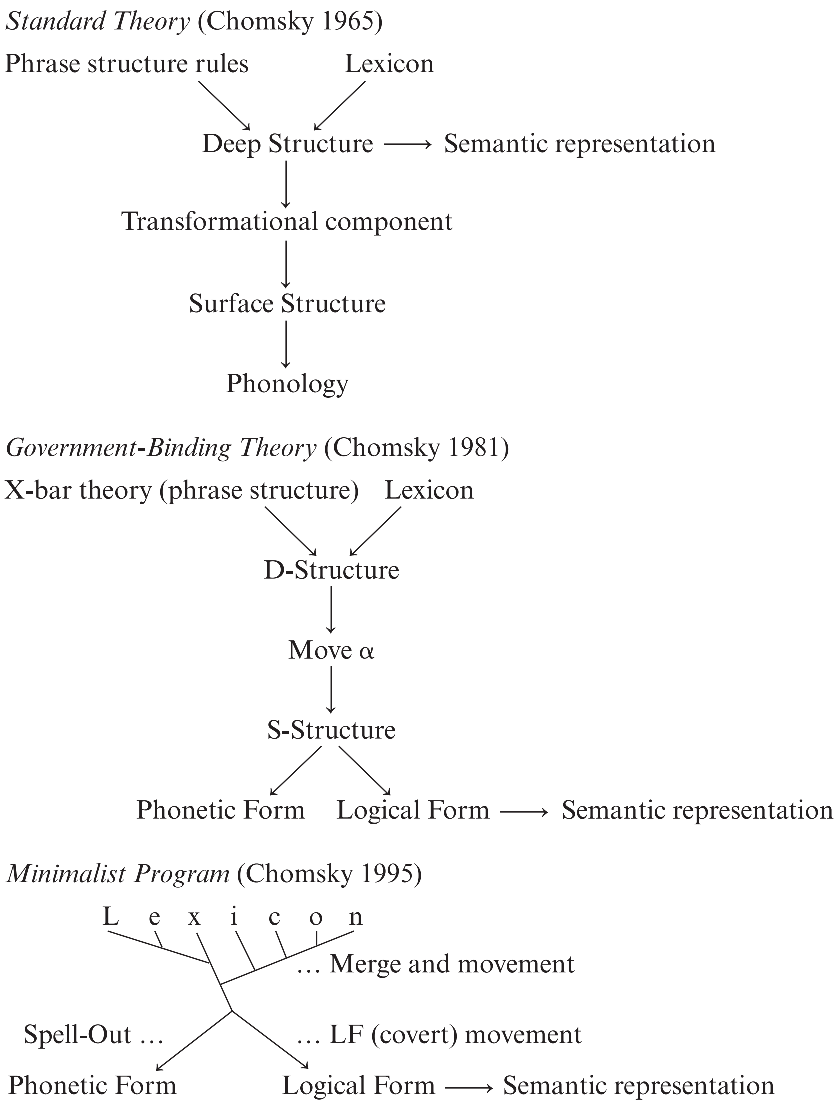
In 1965, this was a perfectly reasonable view. The important issue at that time was to show that *something* in language is generative. Generative syntax had provided powerful new tools, which were yielding copious and striking results. At the time, it looked as though phonology could be treated as a sort of low-level derivative of syntax: the syntax gets the words in the right order, then phonology massages their pronunciation to adjust them to their local environment. As for semantics, virtually nothing was known: the only proposals on the table were the rudimentary treatment by Katz and Fodor (1963) and some promising work by people such as Bierwisch (1967, 1969) and Weinreich (1966). So the state of the theory offered no reason to question the assumption that all combinatorial complexity arises from syntax.

Subsequent shifts in mainstream generative linguistics stressed major differences in outlook. But one thing that remained unchanged was the assumption that syntax is the sole source of combinatoriality. Figure 2.1 diagrams the architecture of components in three major stages of Chomskyan syntactic theory: the so-called Standard Theory (*Aspects*: Chomsky 1965), Principles-and-Parameters (or Government-Binding) Theory (Chomsky 1981), and the Minimalist Program (Chomsky 1995). The arrows denote logical direction of derivation.

The shifts among the theories in figure 2.1 alter the components of syntax and their relation to sound and meaning. What remains constant throughout, though, is that (a) there is an initial stage of derivation in which words or morphemes are combined into syntactic structures; (b) these structures are then altered by various syntactic operations; and (c) certain syntactic structures are shipped off to phonology/phonetics to be pronounced and other syntactic structures are shipped off to "semantic interpretation" to be understood. In short, syntax is the source of all linguistic organization.

I believe that this assumption of *syntactocentrism*—which, I repeat, was never explicitly grounded—was an important mistake at the heart of the field.[4] The correct approach is to regard linguistic structure as the

[4] Some opponents of generative grammar (e.g. some Cognitive Grammarians) have rightly objected to syntactocentrism, but have proposed instead that all properties of language are derivable from meaning. I take this to be equally misguided, for reasons that should be evident as we proceed. See also Culicover and Jackendoff 2005, sec. 1.4.5.



Standard Theory (Chomsky 1965)

Government-Binding Theory (Chomsky 1981)

Minimalist Program (Chomsky 1995)

Figure 2.1
Architecture of mainstream theories over the years

product of a number of parallel but interacting generative capacities—at the very least, one each for phonology, syntax, and semantics. As we will see, elements of such a *parallel architecture* have been implicit in practice in the field for years. What is novel in the present proposal is bringing these practices out into the open, stating them as a foundational principle of linguistic organization, and exploring the large-scale consequences.

2.5 Phonology as an Exemplar of the Parallel Architecture

An unnoticed crack in the assumption of syntactocentrism appeared in the mid to late 1970s, when the theory of phonology underwent a major sea change. Before then, the sound system of language had been regarded essentially as a sequence of speech sounds. Any further structure, such as the division into words, was thought of as simply inherited from syntax. However, beginning with work such as that of Goldsmith (1979) and Liberman and Prince (1977), phonology rapidly came to be thought of as having its own autonomous structure, in fact multiple structures or *tiers*. Figure 2.2 provides a sample, the structure of the phrase *the big apple*; the upper part of figure 1.1 is a larger example. The phonological segments appear at the bottom, as terminal elements of the syllabic tree.

There are several innovations here, already mentioned in section 1.3. First, syllabic structure is viewed as hierarchically organized. At the center of the syllable (notated as *σ*) is a syllabic nucleus (notated *N*), which is

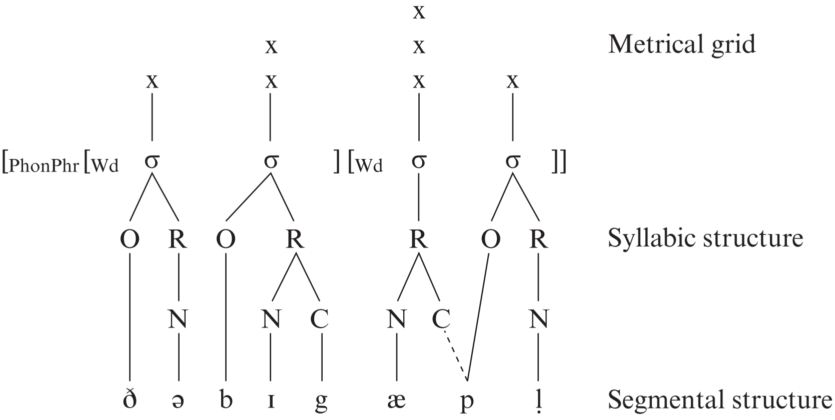


Figure 2.2
Phonological structure of *the big apple*

usually a vowel but sometimes a syllabic consonant such as the *l* in *apple*. The material following the nucleus is the syllabic coda (notated *C*); this groups with the nucleus to form the rhyme (notated *R*), the part involved in rhyming. In turn, the rhyme groups with the syllabic onset (notated *O*) to form the entire syllable. Syllables are grouped together into larger units such as feet and phonological words (here, the bracketing subscripted *Wd*). Notice that in figure 2.2, the word *the* does not constitute a phonological word on its own; it is attached (or cliticized) to the word *big*. Finally, phonological words group into larger units such as phonological phrases (here, the bracketing subscripted *PhonPhr*). Languages differ in their repertoire of admissible nuclei, onsets, and codas, but the basic hierarchical organization and the principles by which strings of segments are divided into syllables are universal. (It should also be mentioned that signed languages have parallel syllabic organization, except that the syllables are built out of manual rather than vocal constituents (Klima and Bellugi 1979; Fischer and Siple 1990).)

These hierarchical structures are not built out of syntactic primitives such as nouns, verbs, and determiners; their units are intrinsically phonological. In addition, the structures, though hierarchical, are not recursive, in that, unlike syntactic structures, they cannot be embedded indefinitely deeply in other structures of the same type.[5] For example, a rhyme cannot be subordinate to a syllable that is in turn subordinate to another rhyme. Thus the principles governing these structures are not derivable from syntactic structures; they are an autonomous system of generative rules.

Next consider the metrical grid in figure 2.2. It is built of nonsyntactic units: its units are *beats*, notated as columns of *x*s. As it is to some degree independent of syllabic structure, it turns out to be an autonomous "tier" of phonological structure. As described in section 1.3, a column with only one *x* is a weak beat, and more *x*s in a column indicate a relatively stronger beat. Each beat is associated with a syllable; the strength of a beat indicates the relative stress on that syllable, so that for example in figure 2.2 the first syllable of *apple* receives maximum stress.

[5] It is important to distinguish two interpretations of "syntactic" here. In the broader sense, every combinatorial system has a syntax: mathematics, computer languages, music, and even phonology and semantics. In the narrower sense of technical linguistics, "syntactic" denotes the organization of units such as NPs, VPs, and prepositions. I am reserving "syntactic" for this narrower sense and using "combinatorial" for the broader sense.

The basic principles of metrical grids are in part autonomous of language: they also appear, for instance, in music (Lerdahl and Jackendoff 1983; Jackendoff and Lerdahl 2006), where they are associated with notes instead of syllables. Metrical grids place a high priority on rhythmicity: an optimum grid presents an alternation of strong and weak beats, as is found in music and in much poetry. On the other hand, the structure of syllables exerts an influence on the associated metrical grid: syllables with heavy rhymes (i.e. containing a coda or a long vowel) "want" to be associated with relatively heavy stress. The stress rules of a language concern the way syllabic structure comes to be associated with a metrical grid; languages differ in this respect in ways that are now quite well understood (e.g. Halle and Idsardi 1995; Kager 1995).

At a larger scale of phonological organization, we find prosodic units over which intonation contours are defined. These are comparable in size to syntactic phrases but do not coincide with them. Here are two examples:

(3) *Syntactic bracketing*
[*Sesame Street*] [is [a production [of [the Children's Television Workshop]]]]

Prosodic bracketing (two pronunciations)
a. [*Sesame Street* is a production of] [the Children's Television Workshop]
b. [*Sesame Street*] [is a production] [of the Children's Television Workshop]

(4) *Syntactic bracketing*
[This] [is [the cat [that chased [the rat [that ate [the cheese]]]]]]

Prosodic bracketing
[This is the cat] [that chased the rat] [that ate the cheese]

The two pronunciations of (3) are both acceptable, and other prosodic bracketings are also possible. However, the choice of prosodic bracketing is not entirely free, since for instance *[Sesame] [Street is a production of the] [Children's Television Workshop]* is an impossible phrasing.

Now notice that the first constituent of (3a) and the second constituent of (3b) do not correspond to any syntactic constituent. We would be hard pressed to know what syntactic label to give to *[Sesame Street is a production of]*. But as an *intonational* constituent it is perfectly fine. Similarly in (4), the syntax is relentlessly right-embedded, but the prosody is flat and perfectly balanced into three parts. Again, the first two constituents of the prosody do not correspond to syntactic constituents of the sentence.

The proper way to deal with this lack of correspondence is to posit a phonological category called Intonation Phrase, which plays a role in the assignment of intonation contours and the distribution of stress (Beckman and Pierrehumbert 1986; Ladd 1996). Intonation Phrases are to some degree correlated with syntax. Their boundaries tend to fall at the beginning of major syntactic constituents; however, their *ends* do not necessarily correlate with the ends of the same syntactic constituents. For instance, the first Intonation Phrase of (3a) begins at the beginning of the sentence, but it does not end at the end of the sentence. At the same time, Intonation Phrases have their own autonomous constraints, in particular a strong preference for rhythmicity and parallelism (as evinced in (4), for example), and a preference for saving the longest prosodic constituent for the end of the sentence.[6]

Another example of mismatch between syntax and phonology comes from contractions such as *I'm* and *star's* (as in *The little star's beside a big star*). These are clearly phonological words, but what is their syntactic category? It is implausible to see them either as noun phrases that incidentally contain a verb or as verbs that incidentally contain a noun. Keeping phonological and syntactic structure separate allows us to say the natural thing: they are phonological words that correspond to two separate syntactic constituents.

(5) Syntactic structure: [$_{\text{NP}}$ I] [$_{\text{V}}$ (a)m] [$_{\text{N}}$ star] [$_{\text{V}}$ (i)s]
Phonological structure: [$_{\text{Wd}}$ I'm] [$_{\text{Wd}}$ star's]

Since every different sentence of the language has a different phonological structure, and since phonological structures cannot be derived from syntax by classical transformational procedures of movement, deletion, and insertion, the usual arguments for combinatoriality lead us to the conclusion that phonological structure is generative: it arises from its own characteristic primitives and principles of combination. Thus we now have two independent generative systems involved in language.

We must next introduce a way to correlate these two systems. This requires a new kind of principle in the grammar, which might be called

[6] Chomsky (1965) analyzes the prosody of an example like (4) as a fact of performance: speakers don't pronounce the sentence in accordance with its syntactic structure. This is about the only way he *can* analyze it, given that he does not have independent principles of intonational constituency at his disposal. Contemporary theory allows us to say (correctly, I believe) that (4) is well-formed both syntactically and prosodically, with a well-formed correspondence between the two structures.

correspondence rules or *interface rules*. These rules (I revert to the standard term "rules" rather than being obsessive about "f-rules") regulate the way the independent structures correspond with each other. For instance, the relation between syllable weight and metrical weight is regulated by an interface rule between syllabic and metrical structure; the relation between syntactic and intonational constituents is regulated by an interface rule between syntactic and prosodic structure.

An important property of interface rules is that they don't "see" every aspect of the structures they are connecting. For instance, the rules that relate syllabic content to metrical grids are insensitive to syllable onsets: (nearly?) universally, stress rules care only about what happens in the rhyme. Similarly, although the connection between syntax and phonology "sees" certain syntactic boundaries, it is insensitive to the depth of syntactic embedding. Moreover, syntactic structure is totally insensitive to the segmental content of the words it is arranging. For instance, there is no syntactic rule that applies only to words that begin with *b*. Thus interface rules implement not isomorphisms between the structures they relate, but only partial homomorphisms.

This is not to say that we should view speakers as thinking up phonological and syntactic structures independently in the hope they can be matched up by the interfaces. That would be the same sort of mistake as thinking that speakers start with the symbol *S* and generate a syntactic tree, finally putting in words so they know what the sentence is about. At the moment, we are not thinking in terms of production; rather, we are stating the principles (of "competence") in terms of which sentences are well-formed. We will get back to how this is related to processing in section 2.10.2.

Now the main point of this section. This view of phonological structure, developed in the late 1970s and almost immediately adopted by phonologists as standard, is deeply subversive of the syntactocentric assumption that all linguistic combinatoriality originates in syntax. Phonological structure proves not to be just a passive hand-me-down derived from low-level syntax: it has its own role in shaping the totality of linguistic structure. But at the time of this shift in phonological theory, no great commotion was made about this most radical aspect of the new approach. Phonologists for the most part were happy to get on with exploring this exciting way of doing things, and for them, the consequences for syntax didn't matter. Syntacticians, for their part, simply found phonology irrelevant to their concerns of constraining movement rules and the like, especially since phonology had now developed its own arcane technical machinery. So neither subdiscipline really took notice; and as the technologies diverged, the relation between syntax and phonology became a no-man's-land (or perhaps only a very-few-man's-land).[7]

2.6 The Syntax-Semantics Interface

I have treated the developments in phonology first because they are less controversial. But in fact the same thing happened in semantics. Over the course of the 1970s and 1980s, several radically different approaches to semantics developed: within linguistics, at least Formal Semantics, which grew out of formal logic (Partee 1976; Larson and Segal 1995; Heim and Kratzer 1998), Cognitive Grammar (Lakoff 1987; Langacker 1987, 1998; Talmy 2000), and Conceptual Semantics (Jackendoff 1983, 1990; Pinker 1989; Pustejovsky 1995), plus approaches within computational linguistics and cognitive psychology. Whatever their differences, all these approaches take meaning to be deeply combinatorial. None of them take the units of semantic structure to be syntactic units in the narrow sense, such as NPs and VPs; rather, the units are intrinsically semantic entities like objects, events, actions, properties, times, and quantifiers. Therefore, whichever semantic theory we choose, it is necessary to grant semantics an independent generative organization, and it is necessary to include in the theory of grammar an interface component that correlates semantic structures with syntactic and phonological structures. In other words, the relation of syntax to semantics is qualitatively parallel to the relation of syntax to phonology. However, apparently none of these schools of thought pointed out the challenge to syntactocentrism—except the Cognitive Grammarians, who mostly went to the other extreme and denied syntax *any* independent role, and who have been steadfastly ignored by mainstream generative linguistics.

[7] As far as I can determine, in all of Chomsky's frequent writings on the character of the human language capacity, there is virtually no reference to post-1975 phonology—much less to the challenge that it presents to his overall syntactocentric view of language. Hauser, Chomsky, and Fitch (2002; also Fitch, Hauser, and Chomsky 2005) advocate treating phonology as part of the "broad language faculty," that is, as an aspect of language shared with other faculties. But the evidence they adduce from animal studies for such a view is quite slim, and it does not address the combinatorial complexity of phonological structure, which has all the flavor of an evolutionary adaptation for precise and efficient communication. See Pinker and Jackendoff 2005, Jackendoff and Pinker 2005 for discussion.

The organization of phonological structure into semi-independent tiers finds a parallel in semantics as well. Linguistic meaning can be readily factored into two independent aspects. On one hand, there is what might be called *propositional structure*: who did what to whom and so on. For instance, in *The bear chased the lion*, there is an event of chasing in which the bear is the chaser and the lion is the "chasee." On the other hand, there is also what is now called *information structure*: the partitioning of the message into old versus new information, topic versus comment, presupposition versus focus, and so forth. We can leave the propositional structure of a sentence intact but change its information structure, by using stress (6a–c) or various focusing constructions (6d–f).

(6) a. The BEAR chased the lion.
    b. The bear chased the LION.
    c. The bear CHASED the lion.
    d. It was the bear that chased the lion.
    e. What the bear did was chase the lion.
    f. What happened to the lion was the bear chased it.

Thus the propositional structure and the information structure are orthogonal dimensions of meaning and can profitably be regarded as autonomous tiers.[8]

Like the interface between syntax and phonology, that between syntax and semantics is not an isomorphism. Some aspects of syntax make no difference in semantics. For instance, the semantic structure of a language is the same whether or not the syntax marks subject-verb agreement, verb-object agreement, or nominative and accusative case. Similarly, the semantic structure of a language does not care whether the syntax calls for the verb to come after the subject (as in English), at the end of the clause (as in Japanese), or second in a main clause and last in a subordinate clause (as in German). As these aspects of syntax are not correlated with or derivable from semantics, the interface component disregards them.

More controversially, some aspects of semantics have little if any systematic effect in syntax. Here are a few well-known candidate phenomena:

[8] In Jackendoff 2002a, chap. 12, I propose a further split of propositional structure into descriptive and referential tiers, an issue too complex for the present context. Chapter 6 discusses a further partitioning of propositional structure into thematic and macrorole tiers, enlarging on a proposal in Jackendoff 1990.

- The syntactic form of a question can be used to elicit information (7a), test someone's knowledge (7b), request an action (7c), or sarcastically express an affirmative answer to a previous question (7d). Thus these choices of illocutionary force are not mapped into syntactic structure.

  (7) a. Where is my hat?
      b. (Now, Billy:) What's the capital of New York?
      c. Would you mind opening the window?
      d. Is the Pope Catholic?

- In example (8a), the interpretation is that Jill jumped multiple times. However, if we change the verb to *sleep*, as in (8b), we don't interpret the sentence as implying multiple acts of sleeping; and if we change *until* to *when*, as in (8c), only a single jump is entailed. Thus the sense of iteration arises neither from any single word in the sentence, nor from the syntactic structure, but from an interaction of *jump*, *until*, and the semantic relation between them.

  (8) a. Jill jumped until the alarm went off.
      b. Jill slept until the alarm went off.
      c. Jill jumped when the alarm went off.

  One standard account of this contrast (Verkuyl 1993; Pustejovsky 1995; Jackendoff 1996c; Talmy 2000) is that the phrase following *until* establishes a temporal bound for an ongoing process. When the verb phrase already denotes an ongoing process, such as sleeping, all is well. But when the verb phrase denotes an action that has a natural temporal ending, such as jumping, then its interpretation is "coerced" into *repeated* action—a type of ongoing process—which in turn can have a temporal bound set on it by *until*. For present purposes, the point is that the sense of repetition arises from semantic combination, without any direct syntactic reflex. (On the other hand, in languages such as American Sign Language that have a grammatical marker of iteration, this marker will have to be used in the translation of (8a).)[9]

[9] This account of coercion is supported by psycholinguistic and neurolinguistic data (Piñango, Zurif, and Jackendoff 1999; Piñango and Zurif 2001): one can detect the additional processing load due to the coercion in (8a) at a time during processing appropriate for semantic integration. Moreover, Broca's aphasics, for whom syntax is impaired but semantics is intact, understand (8a) to involve repetition, whereas Wernicke's aphasics, whose semantics is impaired, are not reliable in detecting coercion.

- In the examples in (9), the “understood” subject of the sentence is not the entity normally denoted by the actual subject (Nunberg 1979).

(9) a. [One waitress says to another]:
The ham sandwich wants another cup of coffee.
[Interpretation: ‘*The person who ordered/is eating* the ham sandwich . . . ’]
b. Chomsky is on the top shelf next to Plato.
[Interpretation: ‘*The book by* Chomsky . . . ’]

Such cases of “reference transfer” contain no syntactic reflex of the italicized parts of the interpretation. One might be tempted to dismiss these phenomena as “mere pragmatics,” hence outside the grammatical system. But this proves impossible, because reference transfer can have indirect grammatical effects. For a clear example, imagine that Richard Nixon went to see the opera *Nixon in China* (yes, a real opera!), and what happened was that

(10) Nixon was astonished to see himself sing a foolish duet with Pat.

The singer of the duet, of course, is the *actor playing* Nixon; thus the interpretation of *himself* involves a reference transfer. However, we cannot felicitously say that what happened next was that

(11) *(Up on stage,) Nixon was astonished to see himself get up and walk out.

That is, a reflexive pronoun referring to the acted character can have the real person as antecedent, but not vice versa. Since the use of reflexive pronouns is central to grammar, reference transfer cannot be seen as “extragrammatical.” Moreover, it proves empirically impossible to encode the understood antecedent of the pronoun as a “hidden” noun phrase in the syntax; these cases can be shown to be true mismatches between syntactic and semantic structure (Fauconnier 1985; Jackendoff 1992b; Culicover and Jackendoff 2005).

- One of the more controversial issues within generative grammar has been the syntactic status of quantifier scope. Consider the two interpretations of (12).

(12) Everyone in this room knows at least two languages . . .
a. ‘for instance, John knows English and French, and Sue knows Hebrew and Hausa.’
b. ‘namely Mandarin and Navajo.’

Should there be two different syntactic structures associated with these two interpretations? Chomsky said no (1957) and later yes (1981); Generative Semantics said yes; I am inclined to say no (Jackendoff 1996c; 2002a, chap. 12; Culicover and Jackendoff 2005). The problem with attributing two different syntactic structures is that it requires systematic and drastic distortions of the syntactic tree that never show up in the surface syntax of any language. The problem with attributing only one syntactic structure is that it makes the syntax-semantics interface more complex. The point to be made here is that the scope of quantification may well be a further example of the "dirtiness" of the interface between syntax and semantics; this continues to be an important issue in linguistic theory.

- Consider a dialogue like (13).

(13) A: What did Bill eat for breakfast?
B: Pizza.

B's response is clearly understood as 'Bill ate pizza for breakfast'. In a theory in which semantics is derived from syntax, the response must therefore be derived from some such sentence as *Bill ate pizza for breakfast*, by deleting (or marking as unpronounced) everything but *pizza*. In a parallel theory, the syntax of B's response may be as simple as the NP *pizza*, and its interpretation is arrived at on the basis of the interpretation of A's question. So far the two approaches look equivalent, except that the syntactocentric theory has this extra syntactic structure. But now consider (14).

(14) A: What kind of pizza would you like?
B: How about pepperoni?

Here B's response is understood as expressing a desire for pepperoni pizza, but there exists no literal expansion of B's response, based on A's question, that expresses this grammatically.

(15) *How about you would like pepperoni kind of pizza?
(ungrammatical and inappropriate!)

Thus the syntactocentric approach must posit syntactically dubious manipulations in order to get B's surface output to come out right. By contrast, the parallel architecture requires no such moves. Rather, the relation between A's question and B's response is taken to be mediated by pragmatic strategies (Gricean or relevance-theoretic) that no one would ever consider syntactic, and that can yield even more distant connections, as in (16).

(16) A: How about some lunch?
B: There's a nice Italian place around the corner.

In each of these cases, a syntactocentric theory is forced to derive the semantic distinctions from syntactic distinctions at some covert level of syntax. This requirement of "interface uniformity" forces syntactic theory into artificial solutions such as empty syntactic structure and elaborate movement, which have no independent motivation beyond providing grist for the semantics (Culicover and Jackendoff 2005). On the other hand, if the semantics is treated as independent from syntax but correlated with it, it is possible to permit a less than perfect correlation; it is then an empirical issue to determine how close the match is. The two approaches lead to quite different solutions; they are far from notational variants. The differences raise empirical issues that have been under exploration for decades (though not among practitioners of the mainstream framework), and they are far from settled.

If we abandon syntactocentrism, it is logically possible that there are aspects of semantics that have no effect on syntax but *do* have an effect on phonology through a direct phonology-semantics interface. Such a treatment is attractive for the correlation between prosody and information structure. For instance, the differences among (6a–c) do not show up in syntax at all—only in the stress and intonation in phonology and in the focus-presupposition relations in semantics. In a syntactocentric theory, one is forced to generate these sentences with a dummy syntactic element [+Focus], which serves only to correlate phonology and meaning and does not affect word order or inflection. (Such was the approach in Jackendoff 1972, for instance.) But such an element does no work in syntax per se; it exists only in order to account for the correlation between phonology and semantics. By introducing a direct phonology-semantics interface sensitive to this correlation, we can account for it with less extra machinery; but of course this requires us to abandon syntactocentrism.

2.7 The Outcome: Parallel Architecture

The argument so far is that theoretical thinking in both phonology and semantics has proceeded in practice as though these structures are due to independent generative capacities. What has attracted far less notice among syntacticians, phonologists, and semanticists alike is that such an organization logically requires the grammar to contain interface components that correlate the independent structures. Carrying this observa-

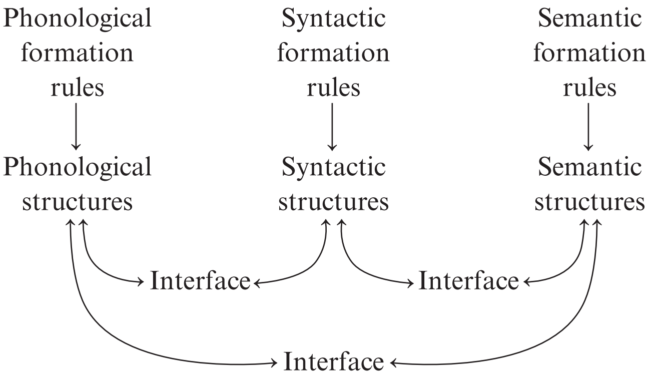


Figure 2.3
The parallel architecture

tion through the entire architecture of grammar, we arrive at an overall picture like figure 2.3, the parallel architecture: the grammar contains multiple sets of formation rules (the "generative" components), each determining its own characteristic type of structure, and the structures are linked or correlated by interface components.

In the syntactocentric architecture, a sentence is well-formed when its initial syntactic tree is well-formed, and all the steps of derivation from this to phonology and semantics are well-formed. In the parallel architecture, a sentence is well-formed when all three of its structures—phonological, syntactic, and semantic—are independently well-formed and a well-formed correspondence among them has been established by the interfaces.

One of the primary interface rules between phonology and syntax is that the linear order of units in phonology corresponds to the linear order of the corresponding units in syntax. One of the primary interface rules between syntax and semantics is that a syntactic head (such as a verb, noun, adjective, or preposition) corresponds to a semantic function and that the syntactic arguments of the head (subject, object, etc.) correspond to the arguments of the semantic function. The consequence of these two primary interface principles is that for the most part, syntax has the linear order of phonology but the embedding structure of semantics.

An illustration of some of these properties of the parallel architecture appears in figure 2.4, the structure of the phrase *the cats* (again, figure 1.1 provides a more complex example). The three independent structures

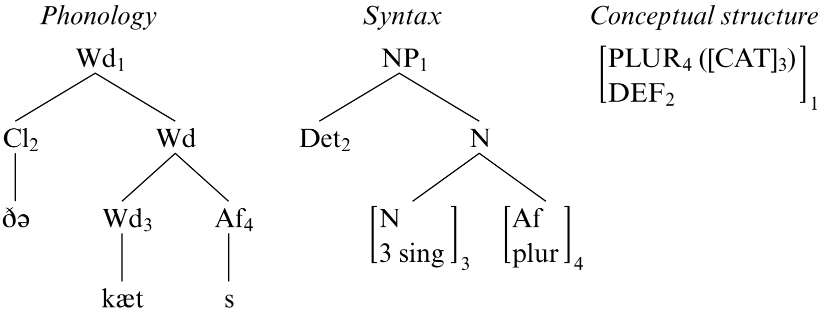


Figure 2.4
The structure of *the cats* in the parallel architecture

are displayed side by side;[10] the subscripting indicates the connections established by the interfaces between the parts of the three structures. For example, the clitic pronounced *ðə* is coindexed with the determiner in the syntax and with the definiteness feature in semantics. Notice that the lowest nodes in the syntactic tree are syntactic features, not the customary notation *the cat-s*. The reasons for this are explained in the next section.

The overall architecture laid out in figure 2.3 provides a model within which many different theories of grammar can be embedded and compared. For instance, figure 2.3 does not dictate whether the syntactic formation rules are along the lines of transformational grammar, the Minimalist Program, Head-Driven Phrase Structure Grammar, or many other alternatives. The syntactocentric framework is a version of figure 2.3 in which the phonological and semantic formation rules are null, so that everything in phonological and semantic structures is determined only by their interfaces with syntax. The framework favored by many in Cognitive Grammar minimizes or even eliminates the syntactic formation rules, so that syntax is determined entirely by meaning.

The organization into parallel generative components is not new here. In addition to the innovations in phonology discussed in section 2.3, a number of proposals within syntax can be mentioned. Lexical-Functional Grammar divides syntax into two parallel tiers, c-structure and f-structure; Autolexical Syntax (Sadock 1991) has a different division

[10] As in figure 1.1, I have used the conceptual structure notation of Jackendoff 1983, 1990, 2002a for the semantics; readers invested in other frameworks should feel free to substitute their own notations.

of syntax into morphosyntax and phrasal syntax; Role and Reference Grammar has, in addition to a morphosyntax/phrasal syntax division, the propositional/information tier division in semantics, with interfaces going every which way among the tiers. In other words, various elements of this architecture are widely present in the literature. What is relatively novel here is recognizing that this organization runs through the entire grammar, from phonology through semantics (in some respects echoing Lamb's (1966) proposals for Stratificational Grammar). As we will see in section 2.10.1, this organization extends further into the rest of the mind as well.

It might well be argued that the standard syntactocentric framework has served the field well for 50 years. Why should anyone want to give it up? A number of replies are possible. First, no one has ever argued *for* the syntactocentric model. In *Aspects*, it was very explicitly only an assumption, which quickly hardened into dogma and then became part of the unstated background. By contrast, the parallel architecture now *has* been argued for, in part on the basis of well-established results in phonology and semantics that have on the whole simply been ignored by the proponents of syntactocentric architectures (or brushed away with phrases like "adopting alternatives that have been proposed would not materially modify the ensuing discussion" (Hauser, Chomsky, and Fitch 2002, 1571)).

A second reason for abandoning syntactocentrism is that, in order to support interpretation and inference properly, semantic structure must be formally far richer than surface syntax. As the examples in the last section illustrate, a syntactocentric architecture requires all semantic combinatoriality (or at least all phrasal semantic combinatoriality) to be read off of syntactic structure. Therefore syntax requires covert representations that reproduce all the complexity of semantic structure. As the theory is extended to account for more and more semantic phenomena, the covert aspects of syntax must be expanded accordingly.

Such expansion has taken place twice in the history of generative grammar. Under the theory of Generative Semantics (see section 2.3), the Deep Structure of Chomsky's (1965) so-called Standard Theory gradually morphed into a bloated tree, full of abstract nodes and requiring copious movement to derive surface forms. Strikingly similar lines of analysis have emerged in Principles-and-Parameters Theory (Chomsky 1981) and its successor, the Minimalist Program (Chomsky 1995): there has been an enormous expansion of abstract structure, driven largely by the need to derive a level of representation, Logical Form, that provides a full and uniform interface to semantics that is independent of the choice of language. The degree of complexity can be appreciated by noting that (in the version of Chomsky 1995) a simple sentence such as *John saw Mary* has a tree with six levels of embedding and four traces (the result of four movement operations). No constituent ends up in the position in which it was introduced, either in the structure fed to phonology or in Logical Form. Culicover and Jackendoff (2005, chap. 3) show that in both Generative Semantics and Principles-and-Parameters Theory, much of this complexity arises from the assumptions of the syntactocentric architecture.

Within a parallel architecture, such an outcome does not arise, because the combinatorial properties of semantics are not first built up in syntax and then transferred to semantics. Rather, they are the product of an independent generative system whose characteristics are attuned to the needs of semantics. To be sure, syntactic structure must be correlated with semantic structure by the interface rules. But this correlation need not be one-to-one; as seen in section 2.6, it can be a good deal more flexible. In particular, the combinatoriality of syntax can in many respects be far simpler than that of semantics. The main constraint is that syntax must be rich enough to account for the mapping between meaning and sound. Under such an approach, worked out in detail in Culicover and Jackendoff 2005, much of the abstractness of mainstream syntax can be dispensed with—including most if not all movement rules and null elements—in favor of a fairly flat "what you see is what you get" articulation of syntactic structure. In particular, standard alternations such as the passive and *wh*-fronting are accounted for not in terms of movement but in terms of special mechanisms in the syntax-semantics interface. This is in fact the style of syntax found in many of the constraint-based syntactic frameworks such as Head-Driven Phrase Structure Grammar and Construction Grammar and assumed by most research in psycholinguistics; the mechanisms have been understood for over 20 years in the Head-Driven Phrase Structure Grammar tradition.

At the same time, syntax does not wither away entirely (as critics of generative grammar would often like). The syntax of a language still has to say where the verb goes, whether the verb agrees with the subject, how to form relative clauses and questions, and so on. The differences among languages in these respects are not predictable from semantics, and children have to learn them.

Nor do the issues of UG disappear if we abandon syntactocentrism. In the parallel architecture, the issues of acquisition and innateness are exactly the same—namely, how do children acquire the grammar of their native language on the basis of environmental evidence? A difference is that the sorts of questions that often arise concern the balance of power among components. We are not forced invariably to ask, what do we have to add to *syntax* to account for such-and-such a phenomenon? Rather, as seen in the examples of the previous section, we find ourselves asking, in which component does this phenomenon belong? Is it a fact of syntax, of semantics, or of the interfaces? And to what extent is it realistic to attribute such a bias to the child learning the language?

Finally, consider the connection of linguistic structure to the rest of the theory of the mind/brain. On the face of it (at least in my opinion), one should favor approaches that permit theoretical integration. Section 2.10 will show four ways that the parallel architecture invites such integration but the syntactocentric theory does not.

2.8 Another Fundamental Mistake: The Lexicon/Grammar Distinction

Every theory of language has to take a word to be a complex of phonological, syntactic, and semantic structures; commonly, the store of words is called the *lexicon*. Explicitly following traditional grammar (e.g. Bloomfield 1933) as well as traditional formal logic (e.g. Carnap 1939), *Aspects* treats the lexicon as a component of language distinct from the rules of grammar. Words are taken to be the locus of irregularity in language, while rules of grammar encode all the regularities. Words get into sentences by being inserted into syntactic trees at the beginning of a syntactic derivation, at the point when syntactic trees are being built and before trees begin to be manipulated and fed to phonology and semantics. This view of the lexicon is retained through all the syntactocentric architectures in figure 2.1. But while it was altogether plausible in the context of early work in generative grammar, I believe that subsequent developments reveal it as another major mistake that has remained in the background as unquestionable dogma within the mainstream school of thought.[11] By contrast, over the past 20 years, the lexicon/grammar distinction has come to be rejected by a variety of alternative frameworks, especially Construction Grammar, Head-Driven Phrase Structure

[11] There has admittedly been some experimentation with ideas like "going back to the lexicon later" for phonological information (as in Distributed Morphology (Halle and Marantz 1993)). Still, the word-by-word assumption is basically preserved.

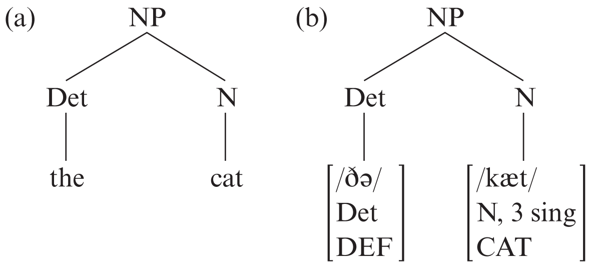


Figure 2.5
Traditional notation for *the cat* (a) and what the traditional notation abbreviates (b)

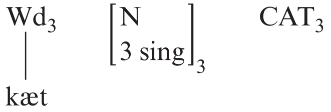


Figure 2.6
The structure of the word *cat*

Grammar, Cognitive Grammar, and Lexical-Functional Grammar, as well as the parallel architecture framework proposed here (Culicover 1999; Jackendoff 2002a; Culicover and Jackendoff 2005).

To see why the standard view of how words get into sentences is problematic, consider the meaning of the traditional notation for trees in figure 2.5a. This is intended as an abbreviation of figure 2.5b, in which the lexical items are spelled out in full. Under this conception, syntactic derivations carry around all the phonological and semantic features of words, which are totally invisible to syntactic rules and are of use to the grammar only when handed over and "interpreted" by the proper component.

The parallel architecture leads to a different treatment. It insists that each kind of feature belongs only in its own structure. Under this view, the traditional syntactic notation in figure 2.5 is formally incoherent, because it has phonological and semantic features as part of a syntactic structure. Thus it should be formally impossible to insert full lexical items into syntactic structure in the traditional way.

How then do words get into linguistic structures? The answer is that each of the three structures making up a word inhabits its own proper type of structure, and each of them carries with it an index that connects it to the others. So, for example, the word *cat* is notated as in figure 2.6;

its contribution to the larger structure in figure 2.4 (*the cats*) should be evident.

Thus a word is best regarded as a type of interface rule that establishes a partial correspondence among pieces of phonological, syntactic, and semantic structure, such that each piece conforms to the formation rules of its own component. In other words, the language does not consist of a lexicon *plus* rules of grammar. Rather, lexical items are *among* the rules of grammar—very particular rules to be sure, but rules nonetheless.

If the distinction between the two theories were confined to the treatment of words alone, it might not raise serious issues—the theories might be considered notational variants. However, the differences are magnified when we consider lexical items that are not single words, for instance idioms. An idiom like *kick the bucket* has to be listed in the lexicon, since its meaning cannot be derived compositionally from the meanings of its constituent words. The classical treatment of lexical insertion (as well as the more recent Merge of Chomsky 1995) requires individual words to be inserted independently, complete with their meanings. It therefore runs into difficulty when the individual words of an idiom in fact have no independent meaning: the meaning can emerge only when the full VP *kick the bucket* is assembled. This problem has been addressed within syntactocentric theories only perfunctorily. Yet any language contains thousands upon thousands of idioms, a substantial part of the vocabulary; this is not a problem that can be lightly dismissed.

Within the parallel architecture, *kick the bucket* can be treated as a lexically listed VP that is coindexed with phonology in the normal way, but that lacks indices connecting the individual words to semantics: instead, the VP as a whole is coindexed with the semantic structure *DIE*. As a consequence, the individual words *kick*, *the*, and *bucket* do not contribute individually to meaning. This is precisely what an idiom is supposed to be: a stored unit in which the words do not have their normal meaning.

Many idioms have normal syntax conforming to general rules: *kick the bucket* is a VP, *son of a gun* is an NP, *down in the dumps* is a PP, *the jig is up* is a sentence, and so on. But a few—such as *by and large*, *for the most part*, *all of a sudden*—have anomalous syntax. These cannot be inserted into standard trees at all by standard lexical insertion, and the classical theory offers no alternative way to insert them. The parallel architecture, by contrast, can treat them as unusual syntactic structures, stored in memory as a unit, and combined in the usual way with the rest of syntactic structure.[12]

Other idioms have variables for open argument places: *take NP for granted*, *give NP a piece of Pronoun's mind*, *put NP in Pronoun's place*, *the cat's got NP's tongue*, for example. This shows that an idiom cannot be treated as an unstructured asyntactic lump, stuffed into otherwise ordinary syntax. Rather, idioms partake of argument structure just like, say, transitive verbs.

The language also contains noncanonical utterance types like those in (17) and other nonstandard syntactic constructions with varying degrees of productivity like those in (18). Presumably, English is not alone in having a sizable number of these "syntactic nuts" (to use Culicover's (1999) term). They all have to be learned, of course.

(17) a. *PP with NP!*
Off with his head!
Into the trunk with you!
b. *How about X?*
How about a cup of coffee?
How about we have a little talk?
c. *NP+acc Pred?*
What, me worry?
Him in an accident? (Akmajian 1984)
d. *NP and S*
One more beer and I'm leaving. (Culicover 1972)
e. *The more S*
The more I read, the less I understand.
f. *How dare NP VP*
How dare Harry question the orthodoxy!
g. *Far be it from NP to VP*
Far be it from Harry to stick his neck out.

(18) a. *Numbers*
three hundred fifty-five million
one hundred twenty-five thousand, six hundred and thirteen

[12] A major concern (perhaps *the* major concern) about idioms within mainstream theory has been why some idioms permit passive and other displacements (e.g. *The cat was let out of the bag*) and others do not (**The bucket was kicked*); see Jackendoff 1997a, chap. 7, for discussion. However, the concern raised here is logically prior, in that it deals with how idioms can get into sentences *at all.*

b. *Focus reduplication* (Horn 1993; Ghomeshi et al. 2004)
You make the tuna salad, and I'll make the SALAD-salad.
Would you like some wine? Would you like a DRINK-drink?
Do you LIKE-her like her?

c. *N-P-N construction* (Williams 1994)
dollar for dollar
face to face
house by house
month after month

These combine the three difficulties of idioms: impossibility of word-by-word insertion, anomalous syntax, and the presence of variables. Within the parallel architecture (as well as Head-Driven Phrase Structure Grammar and Construction Grammar), they can be listed as irregular pieces of syntax including variables, connected with a meaning. As far as I know, advocates of the standard architecture have not addressed these phenomena at all.

A possible reply from advocates of the standard architecture would appeal to the distinction made by Chomsky (1981) between "core grammar"—the deep regularities of language—and the raffish "periphery," which includes "phenomena that result from historical accident, dialect mixture, personal idiosyncrasies, etc." (Chomsky and Lasnik 1993, 510). Chomsky and Lasnik advocate "putting aside" such phenomena, which presumably include idioms and constructions of the sort in (17)–(18).

This defense is unsatisfactory for several reasons. First, idioms and constructions are not "peripheral" to language on any ordinary understanding of that word. As mentioned above, the number of idioms and constructions that speakers know is of a comparable order of magnitude to the number of words, and the frequency of such constructions in text and conversation is very high.

Second, as may be inferred from the examples already presented, it is impossible to draw a sharp line between "core" and "peripheral" phenomena, because the totally regular phenomena of language shade off gradually into idiosyncrasy, and what may be regular in one language (say causative formation) may be only partially regular in another.

Third, the syntactic nuts use the same mechanisms of phrase structure and argument structure as the "core" phenomena of canonical words and structures. For instance, as already noted, idioms such as *take NP for granted* require arguments, just like ordinary transitive verbs. More problematic, there are idioms that can override even the most basic mechanisms of recursive combination that are assumed to be at the heart of the language faculty. For instance, consider the English VP constructions in (19), discussed in Jackendoff 1990, 1997b and Goldberg 1995.

(19) a. He sang/drank/slept/laughed *his head off.*
(V Pronoun's head off = 'V excessively')
b. Bill belched/lurched/joked/laughed *his way out of the meeting.*
(V Pronoun's way PP = 'go PP while/by V-ing')
c. Sara slept/drank/sang/laughed *the whole afternoon away.*
(V NP away = 'spend NP amount of time V-ing')
d. The trolley squealed/rumbled *around the corner.*
(V PP = 'go PP, inducing V-ing sound')
e. Bill drank *the pub dry.*
(V NP AP = 'make NP AP by V-ing')

The italicized complements in these examples are not determined by the verb, as would happen in standard situations. Indeed, these constructions preclude the verb's taking its own object: *He drank (*scotch) his head off, Sara drank (*scotch) the whole afternoon away.* As it were, the construction co-opts the object position for its own nefarious purposes. Goldberg and I argue that these constructions are idioms with VP structure, in which the verb functions as an argument rather than playing its usual role as semantic head. Hence these "peripheral" phenomena commandeer the same computational machinery as the "core" phenomena of phrase structure and argument structure; they are not simple, undigested lumps inserted into language by some separate mechanism. They receive straightforward analysis within the parallel architecture's treatment of the lexicon, but present a severe (and again unaddressed) challenge to the classical architecture.

The conclusion from these widespread phenomena is that human memory must store linguistic expressions of all sizes, from individual morphemes to full idiomatic sentences (such as *The jig is up*). Furthermore, these expressions fall along a continuum of generality, defined by the number and range of variables they contain. At one extreme are wordlike constants such as *dog*, with no variables to be filled. Moving along the continuum, we find mixtures of idiosyncratic content and open variables in idioms like *how dare NP VP* and *take NP for granted*. Still more general are the argument structures of individual predicates such as *dismantle NP* and *put NP PP*. Finally, at the other extreme are rulelike expressions consisting only of very general variables such as *VP → V (NP)*, which like all the others specify possible structures in the language. As a consequence, the formal distinction between lexical items and rules of grammar vanishes.

Within such a theory of the lexicon, what is the difference between a word and a rule? Both are pieces of structure stored in long-term memory. What makes something specifically a rule is that it has variables as part of its structure. (20) illustrates the smooth transition from an idiosyncratic structure to very general principles of language (variables within these structures are notated in italics).

(20) a. VP idiom with no variables: $[_{\text{VP}}[_{\text{V}}\ \text{kick}]\ [_{\text{NP}}[_{\text{Det}}\ \text{the}]\ [_{\text{N}}\ \text{bucket}]]]$
b. VP idioms with variable: $[_{\text{VP}}[_{\text{V}}\ \text{take}]\ \textit{NP}\ [_{\text{PP}}[_{\text{P}}\ \text{to}]\ [_{\text{NP}}\ \text{task}]]]$
$[_{\text{VP}}\ \textit{V}\ [_{\text{VP}}\ \textit{Pronoun}\text{'s head}]\ [_{\text{Prt}}\ \text{off}]]$
c. Standard verb with subcategorization: $[_{\text{VP}}[_{\text{V}}\ \text{put}]\ \textit{NP PP}]$
d. VP structure with more variables: $[_{\text{VP}}\ \textit{V}\ (\textit{NP})\ (\textit{PP})]$
e. Head parameter for VP: $[_{\text{VP}}\ \textit{V}\ldots]$
f. X-bar theory: $[_{\text{XP}}\ldots \textit{X}\ldots]$

(20a) is a stereotypical idiom: a VP with all the phonological material filled in and a stipulated meaning. The examples in (20b) introduce a variable. *Take NP to task* is an idiom with a direct object to be filled in both in syntax and in interpretation; *V Pronoun's head off* is the constructional idiom illustrated in (19a), in which the verb is a variable and fits into the interpretation of the idiom. (20c) is the subcategorization feature for the verb *put*, requiring NP and PP complements. (20d) is more rulelike. It is composed entirely of variables; in fact, it is a notational variant of a standard phrase structure rule for VP. (20e) bleeds more structure out, leaving only the stipulation that the verb is initial in the VP—in effect, the setting of the head parameter for the English VP. Finally, (20f) says that an XP has an X somewhere within it; this is a way of stating X-bar theory, the hypothesis that a phrase has a head of the appropriate category.

(20) illustrates the larger point that the "core" principles of phrase structure are general schemas along the lines of (20e,f), whereas more idiosyncratic rules and fully specified items are usually specializations of these schemas. That is, these items fall into an *inheritance hierarchy* (to use a term common in constraint-based frameworks such as Head-Driven Phrase Structure Grammar and Construction Grammar): (20a–c) are special cases of (20d), (20d) is a special case of (20e), and (20e) is a special case of (20f). On the other hand, there can also be idiosyncratic rules that are not specializations of more general principles, for instance the

N-P-N schema (e.g. *day after day*, *month by month*), which is not an instance of X-bar theory.

Inheritance hierarchies are not specific to language; they are more broadly useful for characterizing knowledge of nonlinguistic categories (e.g. birds and mammals are special cases of animals; cats are special cases of mammals; my late cat Peanut is a special case of cats). Thus this fashion of arranging items in memory comes to the language capacity "for free." What is specific to the language capacity is the sorts of elements that are arranged in inheritance hierarchies: complexes of phonological, syntactic, and conceptual structure that constitute the words and rules of a language.

I must stress that no one in modern linguistic theory has ever argued for a strict lexicon/grammar distinction: it is simply an assumption, carried over from traditional grammar, that roughly accords with common sense. All the alternative frameworks mentioned above have arrived at the dissolution of this distinction through examination of idioms and constructions, which have played virtually no role in mainstream theoretical discussion. If this conclusion is correct, it is a deep and important insight that forces a major rethinking of our vision of language. Such a rethinking is impossible within the assumptions of the syntactocentric architecture.

2.9 The Words-and-Rules Controversy

The parallel architecture view of the lexicon has consequences for the treatment of morphology as well. Irregular morphology constitutes a sort of converse of syntactic idioms. Consider something like the irregular plural *feet*, which must be learned as a distinctive item. It has to be listed syntactically as a plural noun, and the two syntactic parts are coindexed in the normal way to semantics: the word denotes multiple entities of the type *FOOT*. However, the syntactic parts are not connected in normal fashion to phonology; rather, the whole syntactic complex is coindexed with the undifferentiated lump *feet* in phonology, as in figure 2.7.

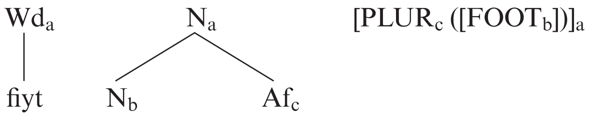


Figure 2.7
The structure of the word *feet*

Notice by contrast how the *regular* plural is encoded in figure 2.4. The regular plural consists of a piece of meaning (namely plurality) plus a piece of syntax (namely an affix attached to nouns) plus a piece of phonology (namely a suffix *s* or *z* or *əz*, the choice determined contextually). That is, the regular plural has all the same parts as a word, and it determines a connection between them. Within the parallel architecture's approach to the lexicon, we can notate this affix as a lexical item along the lines of figure 2.8. (As in (20), the italicized bits denote contextual features that determine how this item is combined with its environment.) The contribution of this item to the overall structure in figure 2.4 is entirely parallel to the contribution of the word *cat*.

This view of regular morphology puts a new and unexpected spin on the by now hoary "words-versus-rules" controversy (e.g. Rumelhart and McClelland 1986; Elman et al. 1996; Pinker 1999). Traditionally, the issue is taken to be this:

- Everyone agrees that irregular plural nouns like *feet* have to be listed in the lexicon. Are regular plurals all listed as well, or is there a separate *rule* for the regular cases that says "To form the plural of a noun, add -*z*"? And, therefore, when children learn to form regular plurals, are they learning something qualitatively different from learning the rough-and-ready generalizations among irregular plurals?

On the present view, words *are* rules—interface rules that help connect phonological, syntactic, and semantic structures. And figure 2.8, the "rule" for the regular English plural affix, is qualitatively no different. Its contextual features are parallel to those of, say, transitive verbs. It combines with nouns the same way a transitive verb combines with its object. Thus the formation of regular plurals is an instance of ordinary combinatoriality. In this approach, the issue comes to be restated like this:

- Are regular plurals all listed in the lexicon, or is there a separate *lexical item* that encodes the regular affix, which combines with any singular

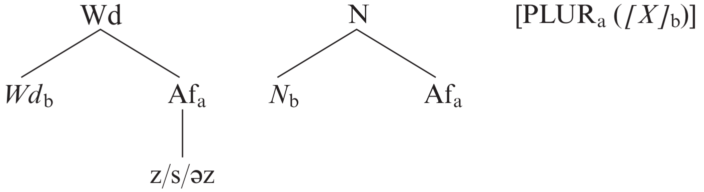


Figure 2.8
The English regular plural as a lexical item

> noun to form a plural noun? And, therefore, when children learn to form regular plurals, are they learning this new lexical item by extracting it as a regularity from the contexts in which it appears—in the same way they extract verbs from the phrasal contexts in which they appear?

I submit that even to the most committed of connectionists, this latter way of framing the question can hardly be objectionable.

In fact, I think the advocates of rules, such as Pinker, have not made the case nearly as strong as it can be. The connectionist argument has been to this effect: we can make a device that learns all English past tenses without making use of a rule, and we can find evidence from acquisition and processing that supports this account. The best version of the anticonnectionist argument has been this: connectionist modeling offers important innovations over standard models of language in dealing with case-by-case learning and analogy for the irregular past tenses; but—you still need rule learning to account for children's acquisition of regular past tense, and we can find evidence from acquisition and processing that supports *this* account. The problem is that the debate has often been framed as though only the past tense were at issue, while the subtext behind the connectionist position is that if *this* can be learned without rules, then it is a good bet that the rest of language can be too.

But not only the past tense is at stake. To deal with the whole of language, it is necessary to account for the creative formation of things like verb phrases and relative clauses—linguistic structures that cannot be listed in the lexicon. On the present view, the way that the regular past tense affix combines with verbs is exactly like the way nouns combine with relative clauses and the way noun phrases combine with verbs and the way subordinate clauses combine with main clauses: it is just another case of free combinatoriality. In the two decades since the original connectionist past tense model was published, there has been no demonstration that the model scales up to acquisition of the full free combinatoriality of language—the issue that grounds generative linguistics.

At the same time, notice that within the parallel architecture, the terms of the dispute become far less contentious. The regular past tense is no longer a qualitatively different phenomenon from words: words are a type of rule, and the posited regular past tense morpheme (figure 2.8) is in the relevant respects just like a word. It differs only in that it is grammatically smaller and requires a word as its grammatical host. So the issue is only whether there is such a separate lexical item, not whether there are two wholly different kinds of linguistic animal, words and rules. Thus in the end the fate of the past tense doesn't seem like such a big deal.

There remains the question of how items stored in long-term memory are assembled into larger composites in working memory. Following approaches such as Head-Driven Phrase Structure Grammar, Lexical-Functional Grammar, and Construction Grammar, the fundamental combinatorial device is taken to be an operation called unification (Shieber 1986). Unification is a sort of Boolean union on feature structures: unifying A and B results in a composite that shares all common features of A and B and preserves all distinct features of A and B. (21) gives two examples:

(21) a. Unification of [V, +past] and [V, 3 sing] = [V, +past, 3 sing]
b. Unification of [$_{\mathrm{VP}}$ V NP] and [V, +past] = [$_{\mathrm{VP}}$[V, +past] NP]

In particular, a variable is instantiated by unifying it with a fully specified structure.

The larger design of language then looks as follows. A typical word is an association of a piece of phonological structure, a piece of syntactic structure, and a piece of meaning. What distinguishes true language from just collections of uttered words is that the semantic relations among the words are conveyed by syntactic and morphological structure. Productivity and compositionality are implemented by the instantiation of variables in stored structures through the process of unification, which applies in phonology, syntax, and semantics. Syntactic rules or principles are regarded as general constructions with maximally unrestricted variables, sometimes but not always bleached of meaning.

In this view, syntax is the solution to a basic design problem: semantic relations are recursive and multidimensional but have to be expressed in a linear string. In particular, propositional structure (who did what to whom) is orthogonal to referential dependencies such as scope of quantification, and both are partly orthogonal to information structure (new vs. old information, topic/focus/common ground). Syntax has to multiplex these conflicting dimensions of structure into a single output representation. Within a language, the result can be alternative word orders: *I saw that movie* (movie is what is seen, presented as new information) versus *That movie, I saw* (movie still is what is seen, but is now presented as the topic). Crosslinguistically, the different dimensions may be simultaneously conveyed using grammatical devices such as case, intonation, and word order. The upshot is that, unlike in the syntactocentric architecture, syntax is not to be regarded as the central generative capacity in language, from which all productivity in expression derives. Rather, it is a sophisticated accounting system for marking semantic relations so that they may be conveyed phonologically.

2.10 Four Ways the Parallel Architecture Helps Integrate Linguistics with Cognitive Neuroscience

The parallel architecture may be an intriguing technical alternative to the Chomskyan orthodoxy in linguistics, but is there any reason why it should be of interest to anyone other than linguists? The previous sections may have begun to offer some hints. This section will sketch a little more fully some ways in which the parallel architecture offers opportunities to unify linguistics with the other cognitive sciences.

2.10.1 The Place of the Parallel Framework in the Larger Architecture of the Mind/Brain

To sum up the larger picture: The parallel architecture claims that language is organized into a number of semi-independent combinatorial systems, each of which has its own organizing principles (a special case of the semimodular design sketched in section 1.5). These systems are linked by systems of interface principles. Interface principles establish a correlation between pieces of structure in two (or more) of the combinatorial systems. Some interface principles deal with large-scale and general correspondences such as the parallel between linear order in syntax and in phonology. On the other hand, some of them are extremely specialized, for instance individual words, idioms, "syntactic nuts," and regular affixes. The interface principles as a whole do not implement an isomorphism between the structures they connect. Rather, they implement a partial homomorphism, a "dirty" correspondence in which not all parts of the structures in question are correlated, and in which many-to-many mappings are altogether common.

This conception of the interfaces within language is perfectly in tune with the way linguistic structures connect to the rest of the mind. Consider how phonology interacts with the auditory system in speech perception and with the motor system in speech production. As is well known (to the continuing dismay of computer scientists working on automated speech recognition), the mapping between a frequency analysis of the speech signal and the phonological structure of an utterance is frighteningly complex. In particular, some aspects of the speech signal play no role in phonological structure and must be factored out, for instance the individual timbre of the speaker's voice, the speaker's tone of voice, and the speed of production, not to mention ambient noise. These aspects of the speech signal are put to use for other cognitive purposes, but not for speech. Moreover, even after all these things are factored out from the acoustic signal, still not every part of the phonological structure is predictable from what is left: most prominently, word boundaries are not present as pauses in the signal. Thus the auditory-to-phonological mapping has the same general characteristics as the interfaces inside language: it too establishes a "dirty" correspondence between certain aspects of two disparate mental structures.

Speech production has similar properties. Not every aspect of phonological structure corresponds to an aspect of the motor control involved in operating the vocal tract. In particular, word boundaries do not correspond at all consistently to pauses in production. Conversely, not every aspect of motor control is controlled by phonological structure. For instance, one can talk intelligibly with a pipe in one's mouth, which hugely distorts the motor commands involved in speech without changing the phonological structure a bit. And of course the same muscles in the vocal tract are used for chewing, swallowing, and so on. Without going into more detail, it should be clear that again the same sort of interface is in play here. (Chomsky's cursory discussion of the "sensorimotor interface" within the Minimalist Program (e.g. Chomsky 2002, 158) surprisingly gives no hint of such characteristics, which have been well known for decades.)

Next consider the visual system. Beyond the very early levels of vision, there is little detailed theory of the f-knowledge involved in vision—the necessary levels of representation and so on. (I take Marr 1982 as an attempt to lay out such a theory, but the enterprise has been largely abandoned since Marr's death.) On the other hand, the neuroscience of vision reveals a qualitatively similar picture: numerous independent brain areas, each specializing in a particular aspect of vision such as shape, motion, color, and spatial relations, each interacting with certain others by dedicated pathways, and no area where "it all comes together" to form a full representation of the visual field. This has precisely the flavor of the parallel architecture in linguistics, where the notion of a "sentence" or "phrase" is distributed among several structures, communicating with each other via dedicated interfaces, as shown in figure 2.3.

Finally, as suggested in section 1.5.2, the lexicon, a large collection of learned arbitrary associations between very particular bits of structure, also has parallels in other domains of memory. Examples include a possible "visual vocabulary" of familiar objects and a cross-modal linking of the appearance of foods to their taste. Just as it is arbitrary that the sound /kæt/ denotes a feline animal, so it is arbitrary (from the point of view of the organism) that something that looks like a cucumber tastes the way it does, and organisms learn probably hundreds or thousands of such associations. There are even "ambiguous"-looking foods: think of mashed potatoes and vanilla ice cream. Similarly, chapter 4 argues that one's knowledge of complex actions also involves a nonlinguistic analogue of the lexicon, in particular a linking of the appearance of artifacts (such as spoons) to schemas for the actions appropriate for them.

A syntactocentric architecture, by comparison, shows no resemblance to the rest of the mind/brain. There is no known parallel to a master "computational system" that generates syntactic structures, which in turn determines phonological structures and meanings. This is one reason Chomsky is forced to say that language is "biologically isolated" (e.g. Chomsky 2002). On his view, even the connection of language to speech is markedly different from the connections among the components inside of language.

2.10.2 The Relation of Grammar to Processing

A theory of linguistic competence is supposed simply to define the permissible structures in the language, without saying how those structures are produced in real time. However, as pointed out in section 2.2, a competence theory ought to lend itself to being embedded in a theory of processing: we ought to be able to say how the f-knowledge that constitutes the competence theory is actually put to use.

There turns out to be an inherent structural reason why competence has to be isolated from performance in the syntactocentric view of language. If we flatten out and simplify the syntactocentric architectures in figure 2.1, they all have a logical directionality proceeding outward from syntax in the middle.

(22) *Logical directionality of syntactocentric architecture*
sound ⇐ phonology ⇐ syntax ⇒ meaning

What I mean by logical directionality is that the possible phonological structures and meanings cannot be determined without first determining syntactic structures. Syntacticians may insist that they are being "metaphorical" when they talk about things happening "before" and "after" other things in a derivation; but the *logical* dependence is there nevertheless.

Now contrast this to the logical directionality of language processing: language perception goes consistently from left to right, and language production from right to left.

(23) a. *Logical directionality of language perception*
sound ⇒ phonology ⇒ syntax ⇒ meaning
b. *Logical directionality of language production*
sound ⇐ phonology ⇐ syntax ⇐ meaning

Hence there is no way that the logical directionality in (22) can serve the purposes of both perception and production. Going from syntax to phonology in (22) seems inherently like production—but only *part* of production; and going from syntax to semantics in (22) seems inherently like perception—but only part of it.

The parallel architecture, by contrast, is inherently *non*directional. The "information flow" between sound and meaning is through the sequence of interfaces, each of which is a system of correlations between two structures, not a derivation of one from the other. The correlations can be used in either direction (which is why they are drawn with double-headed arrows in figure 2.3). We can think of speech perception as a process where structures are activated first at the auditory end of the chain of structures, "clamped" by the environmental input. The interfaces propagate activation rightward through the chain, each interface principle creating a partial resonance between the structures it connects. Eventually, the structured activation reaches semantic structure, at which point it can interact with the hearer's understanding of the context to produce the understanding of the heard utterance. Similarly, in speech production, the speaker begins with a thought to convey; that is, meaning is "clamped" by the speaker's communicative intent. Then the interface principles propagate activation leftward through the chain, eventually activating motor control of the vocal tract and producing speech. Crucially, except for the auditory and vocal parts of the chain, the very same structures and the very same interface principles are invoked in perception and production, just in opposite directions.

There is no need in this system for all of one level to be totally processed before activation of the next level sets in. Any activation of a level, no matter how incomplete, if detected by the next interface, will start to propagate to the next level in the chain. Thus processing can be thought of as "incremental" or "opportunistic" rather than rigidly regulated. In addition, since the interfaces are trying to achieve "resonance" (optimal mapping between levels), there is ample room in the processing theory for feedback in processing—semantics affecting syntactic processing in perception, and vice versa in production.

A crucial tenet of this theory, though, is that *the rules of grammar (including the lexicon) are the only source of information flow in language processing*. For example, knowledge of context cannot directly affect phonological processing, because there are no interface rules that directly relate contextual understanding to phonological structure. On the other hand, context can *indirectly* affect phonological processing—via the interfaces linking them through semantics and syntax. The prediction is that such feedback will take effect some time after constraints directly from phonology, because it has to go up the chain of interfaces and down again. On the whole, such a prediction seems consistent with the experimental literature (Levelt 1989; Cutler and Clifton 1999); many details are worked out in Jackendoff 2002a, chap. 7, in particular the relation of long-term memory to working memory during language processing. (See also Jackendoff 2007.)

The role of the lexicon in the processing theory is entirely parallel to its role in the competence theory. Recalling that words are little interface rules, providing partial routes for mapping between sound and meaning, consider the logic of language perception. The auditory system and the interface from audition to phonology produce some string of speech sounds in the hearer's head, and this activates a call to the lexicon: "Do any of you guys in there sound like *this*?" And various items raise their hands (i.e. get activated). At this point, the processor has no way of knowing which of these items is semantically appropriate, because no contact has yet been made with semantics. However, each item over time activates a connection to potential syntactic and semantic structures, which can be integrated with previous words and with context to determine which candidate word makes the most sense in context. This scenario corresponds precisely to the results in lexical access experiments (Swinney 1979; Tanenhaus, Leiman, and Seidenberg 1979), in which every possible sense of a given phonological string is activated at first, later to be pruned down by semantic context.

A parallel story can be told for speech production. The speaker has activated some conceptual structure that he or she wishes to communicate. The first step is to call the lexicon: "Do any of you guys in there mean *this*?" And various items raise their hands. All the lexical retrieval and speech error literature now comes to bear in showing us the flow of information from this point to actual vocal production; for the most part, it proves to correspond nicely to the options made possible by the components of the parallel architecture (Levelt 1989, 1999).

It is significant that the parallel architecture accords words a very active role in determining the structure of sentences, in concurrence with evidence from the psycholinguistic literature. By contrast, the syntactocentric architecture views words as relatively passive: they simply ride around at the bottom of syntactic trees, while the derivational rules of syntax do all the interesting work. Thus, in the area of the lexicon, the syntactocentric framework again makes it hard to connect competence and performance.

The conclusion here is that the parallel architecture permits a far closer relation between competence and performance theories. The rules of the language, including the words, are posited to be precisely what the processing system uses in constructing mappings between sound and meaning. This opens the door for a two-way dialogue between linguistics and psycholinguistics. Linguistics has always dictated the structures that psycholinguistics should be investigating. But now there is the possibility that psycholinguistic experiments may help determine what component of the grammar is responsible for a particular phenomenon. For instance, experiments by Piñango, Zurif, and Jackendoff (1999) and Piñango and Zurif (2001) have shown that aspectual coercion (e.g. the sense of repetition in *Jill jumped until the alarm went off*, example (8a) above) causes a processing load at a point in time that is appropriate to semantic, not syntactic, processing. This result conforms to the theoretical claim that aspectual coercion is a matter of adjusting semantic well-formedness, not a matter of syntactic deletion of an iterative morpheme. In short, the idealization of a competence theory is not a rigid abstraction; rather, it is a convenient methodological move, to be bridged freely when the occasion arises. (For further discussion of the relation between the parallel architecture and various theories of processing, see Jackendoff 2007.)

2.10.3 The Role of Semantics

Another important advantage of the parallel architecture lies in the way it treats the connection of semantics to the rest of the mind/brain. Elsewhere (Jackendoff 1983; 2002a, chaps. 9 and 10), I have advocated that if generative grammar is to truly adopt the mentalist stance, we must apply it to meaning as well as to syntax and phonology. According to the mentalist stance, the basic function of language is to convert thoughts into communicable form; the virtue of human language over other natural communication systems is that the range of messages it can convey is so broad.[13] Each of the indefinitely many sentences of a language conveys a different thought. Since not all these thoughts can be stored in a single head, it is necessary that thoughts be constructed combinatorially. Therefore an important goal for semantic theory is to uncover the combinatorial system underlying human concepts. Such a goal converges with important trends in psychology and philosophy. Moreover, in a sense it reconceives the promise of the 1960s that generative grammar can be used to discover the character of thought.

However, another influential strain in semantics (and the predominant one in Anglo-American philosophy, dating back to Frege (1892) and shared by people as different as David Lewis (1972), Hilary Putnam (1975), John Searle (1980), and Jerry Fodor (1987)), takes semantics to be the study of the connection of language to the *world*. On this view, a proper semantics has to be concerned above all with how the noise /kæt/ is connected with actual cats. How language users make that connection is quite a different issue (and to many semanticists, not of interest). There is no room here for a critique of this view. I have touched on it in section 1.2; and in Jackendoff 2002a, chaps. 9 and 10, I take up the argument in detail. As in section 1.2, the overall conclusion is that even if one eventually aspires to such a "realist" semantics, the enterprise of discovering how language users *grasp* meaning is also worthwhile. I don't care whether you call the latter enterprise "semantics" or "shmenatics" or whatever: it's *this* enterprise whose central issues intercalate naturally with those of generative linguistics, cognitive psychology, and neuroscience. Just to be clear, I will call it *conceptualist semantics*.

Conceptualist semantics requires us to rethink the traditional issue of reference, which takes as its starting point the unshakable intuition that the phrase *my cat* does indeed pick out an individual in the world. In a

[13] Chomsky sets himself apart from common sense here in his oft-repeated claim that language is not "for" communication (e.g. in Chomsky 2005, citing with approval suggestions to this effect by biologists). For reasons unclear to me, he has always seemed to believe that language came into existence primarily as an aid to thought (if there was any reason for it at all). In at least one recent work (Chomsky 2002) and in a discussion at a conference in spring 2002, he has justified this stance on the grounds that most use of language is for inner speech. Surprisingly, then, he has fallen into the trap to be discussed in chapter 3, that of believing that inner speech *is* thought, rather than (as I will argue) the phonological structure corresponding to thought. See Pinker and Jackendoff 2005, Jackendoff and Pinker 2005 for discussion of (our interpretation of) Chomsky's position.

mentalist linguistic theory, the language user's linguistic system connects the phonological string /maykæt/ to the concept of a feline animal, and to the concept of this feline animal's being possessed by the speaker of the phrase. How then does the language user get from there to the actual individual out there in the world? In brief, the answer is that it isn't just language users who have to connect something in their head to a sense of individuals in the world: any organism with a visual system about like ours (e.g. a baby or an ape) has precisely the same problem. The environment acting on the visual system produces some set of activations in the brain, resulting in the organism's experiencing real objects out there. In other words, conceptualist semantics allows us to recognize that the problem of reference is not a problem about *language*, it's at bottom a problem about *perception* and *cognition* that has to be solved by psychology and neuroscience (see the discussion of visual indices in section 3.4). By contrast, conventional realist theories of reference divorce reference from the mind and make no contact whatsoever with research on perception.

In order for the system of meaning to be influenced by perception, of course, there has to be an interface between conceptual/semantic structure and the "upper end" of the perceptual systems, where "the world" (i.e. the perceiver's conceptualization of the physical world) is organized in terms of stable three-dimensional objects that are located in space with respect to the perceiver and each other, and independently of whether they are sensed visually, tactilely, or (in the case of one's own body) proprioceptively. We might call this level of structure *spatial structure*; it was sketched in figure 1.1 as two stars surrounded by a dotted line.

The interface between conceptual/semantic structure and spatial structure too can be shown to have the standard characteristics: it is a partial homomorphism between the quasi-algebraic format in which linguistic meanings are encoded and the quasi-geometric/topological format(s) in which spatial understanding is encoded (Jackendoff 1987, 1996a). Thus at the semantic end of the language faculty, just as at the phonological end, the relation between language and the rest of the mind is of the same general character as the interfaces within the language faculty itself.

Studying the conceptual system as a combinatorial system leads to the same questions about acquisition as studying syntax. How does the child learning language acquire the meanings of all those thousands of words on the basis of experience, both perceptual and linguistic? What perceptual biases and innate structures does the child bring to the task of interpreting the world? Here conceptualist semantics makes contact with a rich literature on word and concept learning and its innate bases (Carey

1985; Baillargeon 1986; Keil 1989; Gleitman and Landau 1994; Spelke et al. 1994; Bloom 2000; to mention only a few parochial examples). Moreover, since humans doubtless share with monkeys and apes at least the parts of the conceptual system dealing with physical space and perhaps some of the parts dealing with social relations and other minds, conceptualist semantics further makes contact with research on primate cognition (Köhler 1927; Premack 1976; Cheney and Seyfarth 1990; Hauser 2000; Povinelli 2000; Tomasello 2000; see also chapter 5).

Again, these are issues that conventional realist semantics cannot address. Nor are they particularly accessible to semantics studied in a syntactocentric linguistic theory, for if the combinatorial properties of semantics were completely attributable to the combinatorial properties of syntax, then nonlinguistic organisms could not have combinatorial thoughts. There are of course important strains of philosophy that have embraced this view, identifying the capability for thought with the capability for overt language. Descartes comes to mind, for instance; but Chomsky too sometimes appears to believe that combinatorial thought is possible only through the agency of language (see Pinker and Jackendoff 2005 for discussion). I think, however, that contemporary cognitive neuroscience has outgrown such a view, and linguistics ought to be able to follow suit gracefully.

Chapters 6–12 return to issues in conceptual structure, its mentalist instantiation, and its evolutionary roots.

2.10.4 Evolution of Language

Let us return to a point from section 2.2. If UG is a human cognitive specialization, it has to be transmitted by genes that have emerged in the course of our evolutionary divergence from the chimpanzees. Of course, actual evidence for the evolution of the language faculty is practically nonexistent. There is some evidence about the evolution of the human vocal tract (Fitch 2000), but the ability to make speech sounds is only one element of language—and of course there are signed languages, which don't involve speech at all. In addition, it has begun to look like many of the relevant mechanisms for auditory perception are already in place in other mammals (Hauser, Chomsky, and Fitch 2002). But the real issue is this: how did the ability to systematically map combinations of concepts into sequences of speech sounds and back again develop in our species, and how did the ability to *learn* such systematic combinatorial mappings develop?

In the absence of evidence, we would like at least to be able to tell a plausible story about the emergence of UG, an important aspect of which is the overall architecture of language. In particular, we would not like to have to explain language through miraculous emergence, given that—as argued by Pinker and Bloom (1990)—it has the hallmarks of being shaped by natural selection. Pinker and Bloom, however, do not offer any concrete proposals about how language evolved. As is well known, Chomsky himself has been notably evasive on the evolution of the language faculty, often seeming to cast aspersions on the theory of natural selection (Newmeyer 1998 collects representative quotations; see also Pinker and Jackendoff 2005). Chomsky is correct that other factors besides natural selection play a role in evolution, for instance structural constraints (no terrestrial animal will develop wheels) and the biochemistry of proteins. Nevertheless, there is nothing in these other factors (so far) that provides any helpful hints on what caused language to emerge.

The logic of the syntactocentric architecture suggests a reason why such evasion has been necessary. The problem lies in providing a route for incremental evolution, such that some primitive version of the faculty could still be useful to the organism. In the syntactocentric architecture, everything depends on syntax. Combinatorial meaning cannot have evolved before syntax, because its structure is totally dependent on the syntactic structure from which it is derived. For the same reason, phonological structure above the word level cannot have evolved before syntax. Thus either the complexity of syntax had to evolve before the complexity of the other components, or all three had to emerge simultaneously.

The second scenario, simultaneous emergence, seems unlikely. But the first, emergence of syntax before phonology and semantics, is hopeless. What would confer an adaptive advantage on a syntactic faculty that just generated meaningless and imperceptible syntactic structures? And what would enable children to acquire such syntactic structures if there were no perceptible output in the environment from which they could learn it (e.g. if everyone else were just *thinking* in language rather than speaking)? We quickly see that, at this very crude level at least, the syntactocentric theory is stuck: there is no logical way to build it incrementally, such that the earlier stages are useful.[14]

The parallel architecture offers a better alternative. The system of concepts that language expresses is an independent generative component in

[14] Hauser, Chomsky, and Fitch (2002) sidestep this argument, taking the position that once syntactic recursion fell into place, all the rest of what was needed for the language capacity was already there for other purposes. Pinker and Jackendoff (2005) take serious issue with their position, for many reasons mentioned here as well.

the mind/brain. Since it is taken to exist to some degree in other primates as well, it also would have existed in our ancestors, prior to language. That is, our ancestors had interesting thoughts, but lacked any way to say them. Meaning therefore would be the first generative component of language to emerge in the course of evolution (Hauser (2000) urges a similar view).

Most speculation about language evolution goes on to say that the earliest stage would have been the symbolic use of simple vocalization, without grammatical organization. Such a stage is logically impossible in the syntactocentric theory, since even single-word utterances have to arise from syntactic structure. But such a stage is quite natural in the parallel architecture: it consists of stored associations of vocalizations and concepts, what might be called a "paleolexicon." Lexical items that can serve on their own as utterances in fact still exist in modern language, for instance *hello*, *yes*, *oops*, *ouch*, and *gadzooks*. The provision for them in language might be viewed as an evolutionary relic of this earliest stage.

Assuming that a larger number of signals would furnish an adaptive advantage, a regimentation of vocalization along the lines of phonological structure would be the next generative component of language to emerge. Phonological organization in effect digitizes vocalizations, making a large vocabulary reliably discriminable and learnable (Hockett 1960; Nowak, Plotkin, and Jansen 2000). (Proto)words at this point would be simply duples of phonological and semantic structure, without syntax.

A next innovation might be concatenating words into larger utterances. However, when words are concatenated, the issue arises of how the meanings of words in a string are related to each other. In a string like *eat apple Fred*, it is pretty clear on pragmatic grounds that Fred is eating the apple and not the reverse. But pragmatics can only go so far: in *chase lion bear*, who is the chaser? However, something as elaborate as English syntax is not entirely necessary to fix this. Even simple functional principles of linear word order actually afford considerable expressive power. For example, the principle "Agent First" would tell us that the lion is chasing the bear and not the reverse. Such a principle can be stated as a direct phonology-to-semantics mapping, relating linear order to semantic function. Indeed, principles like this appear to be widespread in pidgin languages (Bickerton 1981) and in the grammars of speakers who have acquired a second language late in life, after the sensitive period (Klein and Perdue 1997).

Finally, principles like Agent First have their limitations too. One can imagine the capacity for modern syntactic structure evolving *last*, as a way of making more complex semantic relations among the words of an utterances more precisely mappable to linear word order in phonology. That is, syntax comes along in evolution as a refinement, a "supercharger" of a preexisting interface between phonology and semantics. This seems exactly appropriate to its function within the parallel architecture.

In short, the parallel architecture cannot tell us exactly how language evolved—I don't think anything can ever tell us that. But it does afford a far more plausible hypothesis than the syntactocentric architecture (Jackendoff 2002a, chap. 8, develops this story in considerably more detail). Thus the parallel architecture opens the door for linguistics to interact far more fully with evolutionary psychology, yet another desirable connection.

2.11 Conclusions

Putting this all together, the parallel architecture makes it possible both to integrate linguistic theory internally, establishing the proper relations among phonology, syntax, semantics, and the lexicon, and to integrate linguistic theory more comprehensively with the brain and with biology. In addition, by liberating semantics from its syntactic shackles, the parallel architecture makes it possible to develop a fully psychological theory of meaning and its relation to perception. These points of connection were precisely what early generative grammar promised but ultimately couldn't deliver; I have tried to show here why syntactocentrism was a major reason behind this disappointment.

Of course, to propose a new architecture only begins the work. It opens major questions about exactly what components the grammar requires and what interfaces connect them. Vast numbers of phenomena have been studied in the context of the traditional architecture; to what extent can the analyses proposed there be duplicated or even improved upon? In particular, a thorough overhaul of syntactic theory is necessary, in order to overcome decades of accretions motivated solely by syntactocentric assumptions. (As mentioned in sections 2.7 and 2.8, Culicover and Jackendoff 2005 begins to undertake this task.) Perhaps the hardest part of all this will be maintaining a sense of global integration, keeping the subdomains of the field in closer touch than they have recently been.

But linguistics alone cannot sustain the weight of the inquiry. We need all the help we can get from every possible quarter. And in return, one would hope that linguistic theory might be a more fruitful source of evidence and puzzles for other fields. Above all, my aspiration is to encourage the necessary culture of collaboration.

Chapter 3

Conscious and Unconscious Aspects of Language Structure

This chapter presents a brief version of an approach to the relation of language and consciousness, the *Intermediate-Level Hypothesis*. The hypothesis was developed initially in my book *Consciousness and the Computational Mind* (Jackendoff 1987) and extended in "How Language Helps Us Think" (Jackendoff 1996b). At the time I was writing the book, the study of consciousness was still pretty much forbidden territory within psychology and cognitive science. But since then discussion of consciousness has burgeoned, to the extent that a mere linguist can hardly pretend to master the literature. With that limitation in mind, this chapter restates and amplifies the Intermediate-Level Hypothesis, brings out certain aspects in more detail, and relates it to some of the more recent research.

3.1 The State of the Art

The classical mind-body problem, dating back at least to Descartes, is how conscious experience can be associated with a physical body, and in particular with the brain. It seems fair to say that the contemporary consensus is to seek an explanation of conscious experience in terms of the activities of the brain and the brain alone, eschewing any sort of Cartesian dualism in which consciousness is an autonomous realm with its own causal capacities. Within this position, there seems to be some residual (but shrinking) commitment to epiphenomenalism, the idea that consciousness could be a separate realm that rides passively on the back of brain activity. The predominant view is a strict materialism, in which consciousness is taken to be an emergent property of brains that are undergoing certain sorts of activity.[1]

[1] One frequently discussed issue that distinguishes these two views is whether there could be such a thing as "philosophers' zombies," individuals who from the outside are indistinguishable from us but are not conscious. An epiphenomenalist

Although the distinction is not usually made explicit, one could assert the materialist position in either of two ways. The first would be "methodological materialism": let's see how far we can get toward explaining consciousness under materialist assumptions, while potentially leaving open the possibility of an inexplicable residue. The second would be "dogmatic materialism," which would leave no room for anything but materialist explanation. Since we have no scientific tools for any sort of nonmaterialist explanation, the two positions are in practice indistinguishable, and they lead to the same research. The conclusion parallels the one reached in section 1.2 with respect to intentionality, and similar comments pertain.

Of course, materialism goes strongly against folk intuition about the mind, which concurs with Descartes in thinking of the conscious mind as associated with a nonmaterial "soul" or the like (this point is argued in Jackendoff 1992a, Boyer 2001, and Bloom 2004). The soul is taken to be capable of existence independently of the body. It potentially survives the death of the body and makes its way in the world as a ghost or a spirit or ensconced in another body through reincarnation. Chapter 5 will discuss the deep conceptual basis behind this belief, which seems to recur in every culture of the world. Needless to say, most people cherish the idea of being able to survive the death of their bodies, so materialism is more than an "astonishing hypothesis," to use Crick's (1994) term: it is a truly distressing and alienating one. Nevertheless, by now it does seem the only reasonable way to approach consciousness scientifically.

The point of a materialist theory is that for every aspect of conscious experience, there has to be a corresponding aspect of brain activity (though not the other way around: many brain activities are unconscious). Presumably, the correspondence is systematic. It is therefore natural to investigate what Crick and Koch (1990, 1995) have termed the "neural correlates of consciousness" or NCC. And in fact much of the discussion of consciousness in cognitive neuroscience is concerned with this problem, with several wildly divergent hypotheses in currency (see the end of section 3.2).

might believe in the possibility of such individuals, but a materialist could not. As Dennett (1991) observes, by hypothesis we wouldn't be able to tell if a zombie were unconscious. But more strikingly, *zombies themselves* couldn't tell either, because in order to behave in a fashion externally indistinguishable from us, they would have to sincerely believe they were conscious!

I should also acknowledge that there are still some dualist or quasi-dualist holdouts.

The term of art for the form in which consciousness presents itself is *qualia*. Standard examples are the blueness of blue and the hurtfulness of pain—"what these experiences are like." The piquancy of the mind-body problem arises from the fact that the experiences of blueness and pain seem to bear no connection to anything to which we can imagine neural firings giving rise. Chalmers (1997) has termed this the "Hard Problem," the real knot in a materialist explanation of consciousness. Many philosophers (see e.g. Searle 1992; Robinson 1997) concur; but many others (see e.g. Dennett 2001; Churchland and Churchland 2003) argue that the problem is not so hard after all. In any event, the use of the term "neural *correlates* of consciousness" usefully finesses the issue: it is not necessary to solve the "Hard Problem" in order to claim that, for instance, firing of neurons in such-and-such area of the brain occurs in direct correlation with the experience of blueness.

There is no question that this trend in research is exciting and has yielded all sorts of new insights about the workings of the brain. Nevertheless, I find a surprising gap in most of the discussion: there is little description of how experience is actually structured—of how qualia are organized into the conscious field. (Metzinger's (2000) term is "phenomenal content.") In the major survey presented by Koch (2004), and in four recent anthologies on the cognitive neuroscience of consciousness, representing many of the major players in the discipline (Shear 1997; Metzinger 2000; Dehaene 2001; Osaka 2003), there are for instance frequent references to the distinction between the ventral and dorsal systems in vision. The consensus is that the ventral system is more closely correlated with what one consciously sees, while the dorsal system is more closely correlated with how one behaves in the environment (e.g. reaching for objects). But there is virtually no discussion of the character of what one sees, of what visual experience of the world is like, of the details of visual qualia (beyond perhaps color).

One of the more explicit proposals in these anthologies is offered by Kanwisher (2001). She shows that when an experimental subject is placed in a situation of binocular rivalry between a picture of a face and a picture of a house, the fusiform face area (FFA) is active when the subject is aware of the face stimulus, and the parahippocampal place area (PPA) is active when the subject is aware of the house stimulus. But she says nothing about what the FFA does that makes a face look like a face, how different faces differ in appearance, how such differences are reflected in FFA activity, and so on. Far less explicit are those who attribute consciousness to a very general brain function such as metarepresentation or reentry (e.g. Damasio 2000; Edelman and Tononi 2000) or to an even more general physical process such as quantum collapse (Hameroff and Penrose 1997): here there is no discussion at all of the character of experience.

Some authors in these anthologies note the paucity of phenomenal description in this literature (see e.g. Hut and Shepard 1997; Revonsuo 2000), and Varela (1997) advocates a more thorough attention to phenomenology. Churchland and Churchland (2003) relate color qualia to the properties of the color system in the brain (perhaps the only aspect of higher-level vision understood in such detail). Beyond that, the only actual attempt at describing qualia I find among them is this:

> But what is a thing? When we look carefully, then we find that what we considered to be an object appears in our consciousness as a bundle of meanings, draped around sense impressions that are far, far less complete and filled in and filled up than the 'real thing' we feel to be present, three-dimensionally, continuous in time.... Its reality? Nothing but a sense of reality. (Hut and Shepard 1997, 317)

I submit that this does not take us very far.

The other major gap in these works is that they deal almost exclusively with visual experience. There are passing mentions of other modalities: for instance, Edelman and Tononi (2000, 140) mention "sensory percepts, images, thought, inner speech, emotional feelings, and feelings of will, of self, of familiarity, and so on" and divide sensory percepts into those of sight, hearing, touch, olfaction and taste, proprioception, kinesthesia, pleasure, and pain; Baars (2003, 16) notes that "inner speech is a particularly important source of conscious auditory-phonemic events." There is of course a reason for the emphasis on the visual system: this is the one whose neuroscience is by far the best understood. But vision alone is perhaps too limited for an understanding of consciousness, which cuts across modalities.

I'm going to ask here about consciousness of language.

3.2 What Parts of Linguistic Structure Are Conscious?

The motivation behind my 1987 book was this question:

> What aspects of linguistic structure correspond most closely to the character of awareness—as it were, to the qualia when one is experiencing speech?

This question concerns what might now be called the *functional correlates of consciousness*, however they may be realized in the brain.

Recall from chapters 1 and 2 that linguistic structure has three major departments: phonological, syntactic, and semantic/conceptual structure. In light of that organization, my intuitive answer to the question was this:

Hypothesis 1 When one is experiencing language, the forms in awareness—the qualia—most closely mirror phonological structure.

In case this judgment is not intuitively evident, notice first of all that we experience language as perceived sound. We can intuitively divide utterances into words and syllables with ease (children can count syllables on their fingers by three years of age). Without too much training, we can even divide the speech stream pretty well into individual speech sounds; this is an essential part of learning to read. We have pretty good intuitions about stress patterns as well: most people can easily say where the main stress of a word lies, though they may be uncertain about subsidiary stress. Note however that not *all* aspects of phonological structure result in discriminable qualia: the decomposition of speech sounds into distinctive features (figure 1.2) is experientially opaque.

Compared with phonology, people have far less intuition about *syntactic* structure: you have to take a course in grammar to be able to identify syntactic categories and draw tree structures adequately, and, as noted in chapter 2, there is still substantial disagreement among linguistic professionals about many aspects of syntactic structure. When it comes to *meaning*, our conscious acquaintance with the structure is still more degraded. People have intuitions about meaningfulness, synonymy, entailment, and so forth, but they have no intuitions at all about the form in which meaning is encoded. In fact, this is the area of linguistics where the professionals, after years of research, still have not settled some of the most basic questions. For instance, for 30 years I've had a running battle with logicians about the necessity of the constituent labeled "Place" in figure 1.1. Imagine having a similar battle about whether there's an "s" sound in *star*.

To pump your intuitions further, consider a case in which phonological structure is presented to you but syntactic and semantic structure are absent, for example a stretch of nonsense syllables or of a language you don't understand (*ishkaploople pukapi datofendle*). You still have a form present in awareness—there are still qualia. That is, phonological structure is sufficient to lead to awareness, and meaning is not necessary. A qualification: a meaningful utterance does indeed have a different experiential character than a meaningless one. In part (as pointed out in Prinz, forthcoming), this is because a sentence that one understands conjures up associated images, linguistic and nonlinguistic, via its meaning. But these are nevertheless images and reflect the *form* of meaning no more than the original phonological structure. We will return to this difference in the next section.

What about the opposite case, when there is a meaning present but no associated phonological structure? This is exactly what's going on in the tip-of-the-tongue phenomenon: you know what you want to say—you have a meaning—but the word won't come out. Here, as William James (1890) observed, aside from a strong feeling that there is something there, and a sense of frustration (we'll get to these feelings in the next section), there are *no* qualia in awareness—just this yawning gap waiting to be filled. In other words, meaning without phonology leads to an absence of the qualia that give experience a form—or perhaps there is a quale of absence or formlessness.

I conclude that phonology is necessary and sufficient for the presence of linguistic qualia, and meaning is neither necessary nor sufficient. This conclusion of course goes strongly against intuition, for we often speak of "conscious thought." But in fact, if we pay attention to the phenomenology of "conscious thought," we find it most often has the form of linguistic images—"inner speech" or a "voice in the head," a Joycean stream of consciousness. And if we pay attention to our linguistic images, we realize that they too have phonological form: they come in words, with syllables, stress, rhythm, and even intonation. The form of the associated thought, a semantic/conceptual structure that is capable of driving inference, is not at all present in experience.

Of course, other kinds of imagery are also possible in connection with thinking, such as visual, musical, or proprioceptive imagery. And some sorts of thinking, such as how to arrange the furniture in the living room, how to season the soup, and how one's new symphony should go, are better supported by imagery that is not linguistic. But such imagery is not useful for supporting the sort of reasoning we associate with language. For instance, there is no way for nonlinguistic imagery to encode quantification (***All** ravens are black*), conditionality (***If** it rains today, . . .*), nonpresent time reference (*I went to Venice **last month***), questions (***Who** killed Roger Rabbit?*), or any sort of abstract concept such as obligation (*I **have to** finish this chapter*). Images of a black raven, rain falling, my going to Venice, a blank figure killing Roger Rabbit, and my finishing this chapter may occur in connection with saying or hearing these sentences. But they leave out the crucial boldfaced parts that connect the thought into a chain of inference.

Returning to language, notice that we can ask bilingual speakers, "Do you think in English, or in French?" And they can give us an answer. However, meaning—semantic/conceptual structure—does not come "in English or in French": it's universal.[2] Making inferences on the basis of a thought doesn't depend on the language the thought is expressed in. What it means to translate from English to French is to take the semantic/conceptual structure of an English sentence and clothe it in the syntax and phonology of French (including French vocabulary). So "thinking in French" is just attaching French syntax and phonology to a thought that could have been expressed in *any* language—and then experiencing this thought via the associated phonological qualia. In other words:

Hypothesis 2 We are aware of the content of our linguistically expressed thoughts only by virtue of experiencing phonological images associated with them, plus other images that are inferentially nonefficacious.

Just to push this point home a little further, consider signed languages, which express meaning by hand gestures rather than vocally produced phonology. It turns out that speakers of signed languages experience their thinking not as sound images, but as either visual or proprioceptive images of hand movement (Elissa Newport, pers. comm.)—just what my hypothesis would predict.

Now comes an important corollary:

Hypothesis 3 The *form* of thought itself is always unconscious.

This conclusion goes against a very deep prejudice. There is a long tradition of asserting that our thought (or our intelligence) is the highest, most noble part of human beings—it's what distinguishes us from the animals. And we commonly assign consciousness a similar awe-inspiring status (e.g. Damasio 2000: images are "the highest level of biological phenomena"). That is, thought and consciousness are both taken to be the highest level of cognition. So it is only natural to conflate the two, to identify consciousness with thought (Baars 2003, 11: "[Consciousness] is king of the hill: all active mental processes make use of it").

[2] This abstracts away from possible "Whorfian" differences among languages that are a consequence of vocabulary with different overtones and grammatical structure that emphasizes different aspects of conceptualization (e.g. tense/aspect, politeness, path of motion vs. manner of motion). These do not stand very much in the way of translation. See Jackendoff 2002a, 291–293.

According to my story, this is a big mistake. First of all, we now know that animals, especially higher primates, *do* think (Köhler 1927; Byrne and Whiten 1988; Cheney and Seyfarth 1990; Hauser 2000; Povinelli 2000; Tomasello 2000, among many others). Perhaps they don't think with as much precision and scope as we do, but they certainly are not just "driven by instinct" like machines, as Aristotle and Descartes believed. Rather, the way we differ qualitatively from animals is that we have the ability to convert our thoughts into communicable form, via phonological and syntactic structure—that is, we have *language*. Hence, according to my story, it is possible for us to be conscious of our thoughts in a way that is impossible for animals: not through awareness of the thoughts themselves, but through the awareness of phonological structure associated with thoughts, which animals lack. We'll see in section 3.5 how such awareness could feed back and improve the character of thought. But right now, I just want you to notice that my proposal strongly dissociates consciousness from thought, while at the same time accounting for the intuition that we experience ourselves thinking in a language. To sum up this part of the argument:

Hypothesis 4 Our linguistic images provide most of our evidence that we are thinking.

Of course, there is lots of thinking that goes on unconsciously—that does not come accompanied by linguistic or other images. We have a name for such thinking: we call it "intuition" or "creativity," and we sometimes accord it even a deeper respect and awe than conscious thought. (On the other hand, when animals do it, we call it "instinct" and accord it *less* respect than conscious thought!) One of the important lessons we have learned from 50 years of work in artificial intelligence is how much more there is to human thought than is consciously present.[3] The amount of sophisticated thought that goes on without awareness has also been stressed by people like Michael Polanyi (1958), and, with a dif-

[3] In particular, one of the deepest problems discovered by researchers in artificial intelligence is the so-called Frame Problem (McCarthy and Hayes 1969) (the philosophical tradition calls it the "problem of meaning holism" (Fodor and Lepore 1992))—the problem of deciding at any moment which of an indefinitely large number of factors in memory and the environment are relevant to current reasoning. Our brains generally solve this problem seemingly without effort, and certainly without awareness, as we see from the fact that it took until twentieth-century computer science to discover how pervasive it is in every simple action we carry out.

ferent emphasis, by Varela, Thompson, and Rosch (1991). So I think there are a number of traditions of research that indirectly support the dissociation of thought and consciousness, demonstrating that much of intelligence is not conscious at all and reinforcing hypotheses 2–4.

There is another side to the prejudice linking consciousness and intelligence. I am claiming that the functional correlate of consciousness is something as seemingly peripheral as a phonological structure. This too goes deeply against the grain of intuition. How can the contents of consciousness consist of just a string of sounds? This hardly accords with our sense of the overwhelming importance of consciousness. However, notice that the form of my argument is based on a kind of evidence rarely considered in discussions of consciousness: a detailed and independently motivated analysis of the functional organization of an entire mental faculty. There is no comparable functional analysis for the visual system, where most discussions of consciousness focus; if there were, it would undoubtedly be far more complex. So at the moment there is no way to pick out the functional locus of consciousness as definitively in vision as in language (though I'll make some suggestions in section 3.4).

If these observations and analysis are correct, they make it impossible to maintain certain fashionable views of consciousness.

- It is often claimed that consciousness is an executive capacity, overseeing the activity of the modules of mind when they get into difficulty—a view espoused by people as different as William James (1890), Jerome Bruner (1983), Marvin Minsky (1968), John Eccles (Popper and Eccles 1977), and Thomas Clark (1997). Along similar lines, Koch (2004) views consciousness as providing an "executive summary" of the current situation, which can be "sent off to the planning stages of the brain to help decide a future course of action" (p. 233). "The function of consciousness, therefore, is to handle those special situations for which no automatic procedures are available" (p. 318). But phonology is hardly an appropriate vehicle for executive activity: it's just a structured string of sounds. Phonological structure is of no use whatsoever to the planning part of the brain. It's the *meaning* that planning needs.
- Koch also proposes (p. 243) that "qualia are potent symbolic representations of a fiendish amount of simultaneous information associated with any one percept—its meaning." Again, this is completely at odds with the phenomenology of linguistic awareness, in which meaning plays no role.
- Other people (see e.g. Hofstadter 1979; Rosenthal 1986; Edelman and Tononi 2000; Singer 2000; Parvizi and Damasio 2001) claim that

consciousness somehow involves the mind's including a representation of itself or a higher-order thought—it consists of or is produced by thoughts about thinking. Again, phonology is simply not an appropriate vehicle for encoding a theory of one's own mind or one's own thoughts; it's just an encoding of sounds (which encode thought). To be sure, there *are* mechanisms of mind that accomplish these tasks of monitoring and metarepresentation, but they are emphatically not to be identified with consciousness (though see caveats in the next section).

- Bernard Baars's view of the conscious field as a "global workspace" (Baars 1988, 1997, 2003) has been taken up by many other researchers (see e.g. Chalmers 1997; Dehaene and Naccache 2001; Churchland and Churchland 2003; and see quotations in section 1.6). But this does not make the right distinction. Surely thought should be in the global workspace alongside phonology, but thought doesn't produce qualia. I would be inclined to identify the "global workspace" with working memory (at least on my own reading of working memory as a "workbench" or "blackboard"; see chapter 1, note 7). And only a certain part of working memory is responsible for qualia—namely the phonological part. Global Workspace Theory fails to make this crucial cut.
- Other theories of consciousness connect it with some general property of neurons. Hameroff and Penrose (1997) and Stapp (1997), for instance, claim that consciousness arises through quantum activity of the neurons. Alternatively, it arises from the activity of NMDA receptors (Flohr 2000) or some "proto-awareness" connected to the receptive fields of neurons (MacLennan 1997). These theories fare even worse: although certain neural activities may be necessary for one to be conscious rather than out cold, there is no reason why such activities produce qualia when associated with phonology, but not when associated with syntax, thought, the production of saccades, and the regulation of heart rate.

In each case, I find that theorists advocating these positions never even ask the question of the functional correlates of consciousness—and they certainly do not address how linguistic structure is articulated. Thus they fail to examine the phenomenology adequately. They just talk about "conscious language" and "conscious thought" as though they are self-evident. And from there they proceed on the basis of prejudices that identify consciousness with thought and intelligence. (Further references to earlier incarnations of these theories are cited in Jackendoff 1987, especially sec. 14.1.)

3.3 A Second Dimension of Consciousness: Valuation

3.3.1 [±external] and [±self-initiated]

However, phonological structure cannot be the sole source of qualia for the cognition of language. There has to be something more. Consider: hearing someone else speak involves constructing a phonological structure, and so does saying something oneself, and so does having a linguistic image. If phonology were all there was to it, these would all seem the same in experience—and they obviously don't. One might have the impulse to make the distinction among them by supposing that language perception is clear and imagery is fuzzy (this seems to be Koch's (2004) answer, for instance). But this will not work. I can perceive someone calling to me from a distance over the noise of traffic, and the perception is quite attenuated. Conversely, my inner voice can be perfectly clear. So the difference in experience—obviously a crucial one—must be attributed to something else that is not present in any of the structure laid out in figure 1.1.

I propose to introduce a distinction in structure that is separate from the linguistic content per se. This distinction will be indicated by means of abstract features associated with (or bound to) the structure of the percept or image. One of these features, which I will call *[±external]*, might signify the difference between percepts, which are [+external], and images, which are [−external]. Another feature, *[±self-initiated]*, might concern the difference between self-initiated and non-self-initiated experiences. These two features allow four combinations, explicated in (1).

(1) a. [+external, −self-initiated]: perception of someone else speaking
b. [+external, +self-initiated]: hearing one's own voice as one is speaking
c. [−external, +self-initiated]: hearing one's inner voice
d. [−external, −self-initiated]: hearing unbidden voices in one's head

In Jackendoff 1987, I called these features "affects" of the percept or image; in Jackendoff 1996b, I substituted the term "valuation." An alternative (suggested by Dan Sperber) is "epistemic status." Whatever the name, the idea is that these features add a felt character to the entities in experience. To contrast with these *valuation features*, I'll use the term *content features* for the mental structures that are correlated with form (or qualia) in experience; in the case of language, the content features are drawn from phonological structure.

It should be stressed that valuations are not characteristics of one's experience *as a whole*. Rather, they are attached to particular percepts and/

or images. I may be simultaneously listening to someone talk and making nasty comments to myself, without losing track of which is which. The former part of my experience has the valuation features in (1a), the latter has those in (1c).

Notice that valuation features, like content features, are subject to error. For instance, anytime you experience an unbidden voice in your head, it's some sort of illusion, since your own mind/brain has created the image. And the experience of a hallucination, say of a voice outside in the street, has the valuation features of a percept (1a), although again it is actually produced solely by the brain of the perceiver. *Realizing* one is having a hallucination consists of noticing the inconsistency of the character of the experience with the rest of one's knowledge.

It is important eventually to ask what aspects of mental and neural processing give rise to these features in experience. One can imagine, for example, that [±external] is the result of a monitor that detects the presence or absence of activation in peripheral (or sensory) areas of the brain linked to the percept in question. But for the moment this is getting ahead of the game. Rather, I want to concentrate here on two issues that are crucial to the description of conscious experience. First, I want to show that valuation features are a "horizontal" aspect of structure that cuts across the "vertical" domains of language, vision, and so on (where "horizontal" and "vertical" are understood in the sense of section 1.1). Second, I want to explore the space of possible valuation features. This will help us see what is required of any theory of the neural correlates of this aspect of consciousness.

To see how the valuation features apply across faculties, let's consider vision. Seeing something blue and square creates square blue qualia in experience (however they are functionally characterized); and *imagining* something blue and square does too. Moreover, we now think that pretty much the same brain areas are involved (Kosslyn 1996). Yet the visual experiences are clearly not the same. So blueness and squareness and the like cannot be all there is to visual awareness. The same valuation features we used for language are applicable, with one exception: since one cannot produce external visual experiences, combination (b) is absent.[4]

[4] Various people have suggested to me that this feature combination might be realized by watching oneself move, watching something move that one is moving, or seeing something that one has oneself made (say a piece of art). I don't want to rule any of these out in principle, but they are less obvious than the corresponding case of language. Of the three, I find the first most plausible.

(2) a. [+external, −self-initiated]: ordinary visual perception
b. [+external, +self-initiated]: (*impossible*)
c. [−external, +self-initiated]: voluntary visual imagery
d. [−external, −self-initiated]: involuntary visual imagery, unbidden visual images

The same remarks apply to unbidden visual images and to visual hallucinations as to the parallel phenomena in language. And of course dreams are the parade case of error in valuation: they are experienced both visually and verbally as [+external], though they are clearly internal; and many parts of them are experienced as [−self-initiated], though they clearly are a product of one's own mind/brain.

Audition in general, including music, has the same possibilities for valuation features as language. I trust the reader to work them out. (Would the sounds made by an instrument one is playing be [+external, +self-initiated]?)

More interesting is proprioception, the sense of body position and movement. Here a question arises as to how to interpret the feature [external]. I suggest the proper interpretation is that [+external] means 'in my body' and [−external] means 'imagined'. We then find a four-way split similar to that in language:

(3) a. [+external, +self-initiated]: voluntary motion of my body
b. [+external, −self-initiated]: involuntary motion of my body (as in twitches and blinks) or motion of my body caused by some force other than my own
c. [−external, +self-initiated]: voluntary images of body motion or position (as when contemplating a jump across a chasm)
d. [−external, −self-initiated]: involuntary images of body motion[5]

This case is extremely important, because combination (3a), voluntary action, encodes the sense of willed action. There has been a small but insistent literature on the relation of consciousness and the sense of free will, sparked in part by experiments by Libet et al. (1983), to the effect that the sense of will emerges some 300 milliseconds *after* the brain activity that initiates motion; extended discussions include Dennett 1984, 1991, 2003 and Wegner 2002. The sense of will emerges in the present approach as a valuation feature, a "feel" that goes with performing an action, arising from some aspect of how the action is initiated.

[5] I can't recall experiencing this combination, but I can imagine having it.

This proposal seems in accord with the literature. It does leave the theory open to the accusation that free will is an "illusion"—but this is a problem that *any* materialist theory of mind/brain must grapple with. It seems to me (as it does to Dennett) that the issue of free will is ultimately of importance not for its own sake but more deeply for an explication of responsibility, especially moral responsibility. Not to brush this problem aside, but under this interpretation the nature of our free will becomes a question of the organization of human concepts rather than a question of metaphysical truth. Some of the relevant organization will emerge in chapter 8.

Can there be illusions in valuation features for proprioception? Some illusions of proprioception seem to be errors in *content* features. For instance, illusions of limb position and movement (Lackner and Dizio 2000) are parallel to visual illusions where something appears in the wrong place. However, a possible valuation error is phantom limb, where one experiences [+external] proprioception in the absence of a limb. It is hard to decide whether neglect, where a patient denies that a limb is his or her own, should be understood as an absence of content qualia for proprioception of the limb or as some sort of valuation error. We might also be looking for cases where an individual's limb is moved by someone else but the individual claims the movement was voluntary. I don't know of such cases, but I wouldn't be surprised if they were reported.

Another possible interpretation of the feature [external] (or perhaps another feature) is as "part of me" versus "not part of me." In this case, denial of ownership of a limb would result from assigning it the valuation "not part of me." Conversely, the sense of a tool as an extension of one's body (e.g. feeling the ball strike the tennis racket rather than the racket press against the hand) would result from assigning the racket the valuation "part of me." And the strange sense attached to bodily excretions (you readily swallow your saliva in your mouth but would be repulsed by the thought of drinking it out of a glass) would come from its having made the transition in valuation to "not part of me."

The main point for now is that the valuation features [±external] and [±self-initiated] apply to percepts and images in all modalities of awareness, not just to language. They can therefore be thought of as a "horizontal" dimension of awareness, contrasting with the "vertical" dimension that distinguishes the faculties from one another.

The next question that arises is what other valuation features there might be in addition to these two and what they might be used to explain. The criteria for a candidate valuation feature are (a) that it encodes a difference in awareness that is a matter of "feel" rather than form (form is expressed by content features); (b) that it applies to multiple modalities; (c) that (at least to some extent) it combines freely with the other valuation features to provide a range of character in awareness; and (d) that there are some illusions that arise through misattribution. I cannot provide an exhaustive list of such features, but here are some plausible candidates.

3.3.2 [±familiar]

Consider the difference between the following two bits of language:

(4) a. To be or not to be, that is the question.
b. To find a useful standpoint for free will and determinism has been fraught with slippery footings and fear.

The first immediately elicits a feel of familiarity, and the second does not. This is not because of anything systematic in their form; thousands of linguistic expressions are familiar (see Jackendoff 1997a, especially chap. 7 on the "Wheel of Fortune corpus") and of course an unlimited number are not.

Similarly with visual percepts: we can judge thousands of faces and places and objects and paintings familiar (different ones for each person), and of course indefinitely many will be judged unfamiliar. All experiments on recognition memory depend on subjects' making such judgments. Auditory percepts too can bear this valuation: consider the experience of turning on the radio and gradually recognizing the piece of music being played, at first without being able to identify it. Nothing changes in its form—only the "feel" of familiarity associated with it. Proprioceptive familiarity is harder to pick out: I suppose most body motions feel familiar. But consider trying to learn a new motor skill, say putting spin on a tennis serve or playing double-stops on the violin or dancing a tango: here one might well have the sensation of [−familiar] body motions.

The example of the music on the radio shows that familiarity does not always equal ability to identify. This is true as well of faces ("I'm sure I know you, but who are you again?") and linguistic phrases ("Where's that phrase from? Some old car ad?").

This feature can be attributed to a stimulus erroneously, for instance when a subject in an experiment claims to recognize stimuli never shown before or fails to recognize previously experienced stimuli (even when there are experimentally measurable effects of such previous exposure in terms of reaction time, priming, or the like). Another such error is déjà vu, where one has a strong impression of familiarity associated with a situation that one knows one has never experienced before. Perhaps the opposite error occurs when one has the experience of seeing something afresh, "as though altogether new." Finally, there is a delusional state called "Capgras syndrome," in which (say) a husband claims that his wife has been replaced by an imposter who looks just like her (McKay, Langdon, and Coltheart 2005). Thus the appearance is familiar, but the person is not! One interpretation of this syndrome is that the machinery for detecting familiarity through affective response has failed, in the present terms a failure of affective features.

The feature [±familiar] interacts with [±external] in an unsurprising way. Percepts, which are [+external], can come with or without this aura of familiarity, but so can images ([−external]). A familiar image, visual or verbal, is usually called a "memory"; an unfamiliar one is called a "new idea." There is no necessary difference in form, only this feel of familiarity or not. And anybody who has been a scholar for a while has experienced the sickening realization that his or her latest great new idea ([−familiar]) in fact stems from something he or she read some years back—an error of valuation in an image.

The upshot is that familiarity versus novelty meets the criteria to be a valuation feature.

3.3.3 [±affective]

Compare these sentences:

(5) a. The little star's beside a big star. [−affective]
b. I love you, darling. [+affective: valence+]
c. How can you be such a total idiot?! [+affective: valence−]

The relevant difference here is not in the form but in the "feel": (5a) is affectively neutral, while the other two carry affective or emotional coloring. In turn, the latter two differ in that (5b) has a positive affective valence and (5c) a negative one (under standard conditions of utterance). Thus this feature might be thought of as the functional connection between perception and emotion.[6]

The same sorts of coloring are obviously available in visual and nonlinguistic auditory stimuli, and the distinction can be applied to images as

[6] One might want to expand the options here to a more detailed articulation of emotional content. I don't have any criteria at the moment to decide whether this is advisable, and if so, how to do it.

well as to percepts. Proprioceptive sensations too can feel good, feel bad, or be neutral, with pain as the canonical [+affective: valence−] body sensation and orgasm as a prominent case of [+affective: valence+].

It might seem a little odd to speak of errors of valuation with respect to this feature. But people suffering phobias and paranoia, for instance, could be characterized as experiencing certain percepts with an unjustified [+affective: valence−] valuation; various drugs invest one's perceptions with the valuation [+affective: valence+], "seeing everything through rose-colored glasses." And for those individuals who have been left affectively flat by brain damage, everything remains [−affective].

3.3.4 [±meaningful]

As observed earlier, nonsense syllables have phonology-like content qualia, just like ordinary language. The differences between (6a) and (6b) lie in the fact that one "feels" a conceptual structure lying behind (6a) that makes it more than a sequence of sounds. (6c) is an intermediate case: the individual words are meaningful, but the whole is not.

(6) a. The little star's beside a big star. [+meaningful]
b. Ishkaploople pukapi datofendle. [−meaningful]
c. Colorless green ideas sleep furiously. [−meaningful], but individual words are [+meaningful]

One might instead claim that the distinction between (6a) and (6b) can be attributed to the fact that one has visual imagery associated with (6a). But this will not do for linguistic meaningfulness in general; consider the sense of meaningfulness you experience in the rather abstract sentence you are reading now. Moreover, the sense of meaningfulness pertains not just to individual sentences but also to entire discourses. When listening to a lecture or reading an article, we may "lose the train of thought": functionally, this amounts to the valuation [−meaningful] coming to pertain to the global situation, even if the parts are comprehensible.

Vision has a parallel distinction, perhaps most easily illustrated in terms of art. Rembrandt's paintings are [+meaningful], in that their visual patterns can be parsed into individuals and objects in coherent configurations. In this sense, Jackson Pollock's are not; many of Chagall's paintings are piecewise meaningful but globally incoherent. Usually, ordinary visual perception is meaningful. But occasionally, say when we try to make out a dimly lighted scene, it is not: there are content qualia but they don't add up to anything. Visual images can be meaningful or not;

dreams are often only piecewise meaningful. Nonlinguistic auditory perception can also be coherent (e.g. Schubert) or incoherent (John Cage). (It's hard to figure out what experience might constitute meaningless proprioception, though.)[7]

We all have experiences in which something gradually begins to make sense after repeated exposure. Alternatively, suddenly it "clicks into place," for example when a visual stimulus changes from a pattern of splotches into a picture of a Dalmatian. Nothing changes in the form of the stimulus, but we now "know what it is." More radical is the confabulation of schizophrenics and people undergoing drug-induced epiphanies, to whom all kinds of things "make sense"—the whole world is full of wonderful patterns and meanings that no one else can appreciate.

There are by now too many combinations of the valuation features to explore all of them and demonstrate their independence. We've seen, however, that meaningfulness crosses modalities, and that it is independent of whether the experience is external or internal. It is also independent of [±self-initiated], since I can say meaningless things myself as easily as I can experience others' utterances as such (an occupational hazard of being a linguist?).

Turning to combinations with the remaining features, an offbeat combination is [−meaningful] but [+familiar] and [+affective]. One might think that something meaningless couldn't have positive emotional valence. But in fact this seems possible. For instance, it's no secret that many Jews recite the Kaddish, the prayer in memory of the dead, with great emotional attachment, but without the slightest idea what the Aramaic text is about—for them it is just a sequence of nonsense syllables.

Another offbeat combination is the valuation [+meaningful] in the absence of content features. This seems an appropriate way to characterize the tip-of-the-tongue experience: despite the phonological gap, there is the formless sense of a meaning one wishes to convey.

Within the category [+meaningful], a number of subcategories open up, pertaining to what is understood. For example, an utterance or a visual display can be perceived as *ambiguous*—and sometimes this is intentional, as in a pun or double entendre. It might also make sense to treat

[7] In the arts, of course, it is a mistake to conflate meaningfulness in the present sense with artistic merit. Both representational and abstract art can range from great to mediocre; there are great and mediocre pieces of music in both the standard tonal and the atonal traditions; and literature, poetry, and drama often make artistic use of the surreal and the absurd.

funny as a valuation, perhaps a particular crossing of meaningfulness and affectivity.

Perhaps the most important distinction within [+meaningful] pieces of experience, though, is between those to which one is "committed" and those to which one is not. I will use the feature *[±committed]* to embrace a number of cases whose unity is not usually recognized. For the clearest case, consider the experience of a declarative sentence, uttered either by oneself or by someone else. Here the feature [±committed] corresponds closely to the philosophical notion of "propositional attitude," which is usually defined as 'belief, desire, and so on' (and then usually explicated only in terms of belief). What is the difference in the content features of sentences that you believe and sentences that you do not? Obviously, nothing: you may not have a belief one way or the other about *The little star's beside a big star*, until you check out the visual scene, and then you come to believe or disbelieve it. The sentence hasn't changed, only your feeling of conviction about it. This means that the sense of belief in a proposition should be encoded as a valuation feature: it affects experience not through form but through "feel."

I am going to call the valuation feature associated with belief [+committed: valence+]. Then disbelief can be [+committed: valence−]. This leaves the feature value [−committed], which seems to be a nice characterization of *entertaining* a proposition.

Next consider *desiring* that such-and-such be the case: one doesn't know whether it is the case or not, but one feels it would be a good thing if it *were* the case. This situation can be captured with the feature combination [−committed; +affective: valence+]. This account suggests that there ought to be a counterpart with negative affect; perhaps this slot in the paradigm is filled by *dreading* that such-and-such is or will be the case.

But this range of possibilities is not confined to declarative propositions. Chapter 8 will demonstrate that *believing such-and-such a proposition* and *intending to perform such-and-such an action* are close conceptual parallels. I defer details till then, but as a first hint, note that the same words are used for the causative of both: *convince/persuade someone that such-and-such is the case* means 'cause someone to believe such-and-such', and *convince/persuade someone to do such-and-such* means 'cause someone to intend to do such-and-such'. This suggests that the notion of commitment can be extended to actions and that the experience of an intended (but not yet performed) action carries the valuation [+committed: valence+]. Then the experience of just turning an action over in one's mind or *considering* it ("Would this be an interesting thing to do?") would correspond to entertaining a proposition and would carry the valuation [−committed]. I suppose the counterpart of disbelief would be actively *avoiding* an action, or *inhibiting* it; this would have the valuation [+committed: valence−]. (7) sums the situation up so far.

(7) a. *Propositional attitudes*
[+committed: valence+]: believing
[+committed: valence−]: disbelieving
[−committed]: entertaining
[−committed; +affective: valence+]: desiring
[−committed; +affective: valence−]: dreading
b. *Actional attitudes*
[+committed: valence+]: intending
[+committed: valence−]: avoiding
[−committed]: considering
[−committed; +affective: valence+]: desiring
[−committed; +affective: valence−]: dreading

Looking at nonlinguistic modalities, the actional attitudes easily generalize to actual actions, not just sentences about actions. In particular, (3a) proposed valuation features associated with voluntary action: [+external, +self-initiated]. But these are incomplete, in that they do not distinguish between intentional and accidental self-initiated actions. This gap can now be rectified: an accidental action is one to which one is not committed.

(8) *Self-initiated actions*
a. [+external; +self-initiated; +committed: valence+]: intentional action
b. [+external; +self-initiated; −committed]: accidental action

(8) leaves the paradigm incomplete, in that there is nothing associated with the value [+committed: valence−]. Speculatively, this might be realized as actions that one performs against one's will, for example under compulsion.

What about in vision? The canonical valuation that goes with meaningful visual experience is [+committed: valence+]. This captures the sense that "seeing is believing": the world is *really out there*, an absolutely fundamental aspect of consciousness. But one can have visual experiences to which one has a negative commitment, most prominently the images recalled from dreams. Another possible case is the experience of representational paintings of fictional characters: one might be either committed to the nonexistence of these characters (like Sherlock Holmes and Santa Claus) or neutral about their reality.

3.3.5 Pairwise Valuations

All the valuations so far have applied to individual percepts. Very speculatively, I'd like to suggest that there are also valuation features that apply to pairs of experienced entities, yielding a "feel" about a dyad. One prominent candidate might be called *[±same]*, which would be associated with two experienced entities that are the same (or perhaps even just similar) in any modality. A second candidate I'd like to consider is hard to characterize with a single word, but I'll call it *[±connected]*. The entities in a [+connected] dyad are sensed as having some influence on each other, symmetrical (e.g. their motions are yoked) or asymmetrical (e.g. one is acting on or causing the other). Lightning and thunder might be a case of [+connected] entities that cut across modalities. Another case might be spoiled food one has eaten and the subsequent nausea. That is, I'm thinking of [+connected] as a valuation that serves as a phenomenological underpinning for far more highly ramified concepts of correlation and causation; some of the conceptual elaboration will appear in the *macrorole tier* of chapter 6. However, at the moment it's hard enough to motivate and justify the valuation features for single percepts, much less the binary valuations, so I will leave the issue here.

To sum up this long section: The features of experience divide into two major classes, the content features and the valuation features. The content features give experience its form—its "qualia"; in the case of language, these features are drawn from phonological structure. The valuation features give experienced entities their "feel"—their sense of reality or unreality, of familiarity or novelty, of volition, of coherence, and of emotional connection. The content features are arranged in a complex hierarchical structure. The valuation features are relatively simple binary or ternary oppositions, attached to constituents in this hierarchical structure. Both classes of features are necessary to describe the character of experience.

As mentioned earlier, I believe the literature has been negligent in describing the content features of awareness. But so far as I know, the idea of valuation features has played no role at all in the discussion of consciousness. I think valuation features are a vital element of phenomenological description. I hope this section has shown how adding them to our descriptive tools makes it possible to become far clearer about a substantial number of important phenomenological issues.

In addition, the notion of valuation adds a new dimension to the quest for the neural correlates of consciousness. For instance, the sense of familiarity versus novelty presumably arises through a process that correlates a structure in working memory with activity in long-term memory; the sense of self-initiated activity requires comparing an action plan in working memory with the perceived result of activity. Thus here, if anywhere, is where the neural correlates of consciousness involve the brain monitoring its own states and activity. The theories of consciousness that stress its reflexive character are in a sense correct—not with respect to content features (as pointed out at the end of section 3.2), but with respect to valuation features, whose existence virtually nobody has noticed.

3.4 The Role of Attention in Consciousness

Returning to the content features for awareness, why should it be that phonological structure, of all things, is the level that functions as the locus of linguistic awareness? My sense is that this is consistent with the character of awareness in other "vertical" modalities of more ancient evolutionary lineage. In the visual system, for instance, we see directly the front surfaces of stable objects. This is something more structured than, say, the contents of primary visual cortex (V1), which are not organized into coherent regions and which move about with every saccade of the eyes. And what we see directly is less structured than full three-dimensional spatial understanding, which has to include an understanding of the backs of objects, their solidity or hollowness, the forces objects exert on each other, and their affordances for manipulation and autonomous action. Thus the phenomenology of vision strikes me as quite parallel to that in language, where phonological structure is somewhere between a raw acoustic analysis and a full-fledged understanding. This is close to the view arrived at by Crick and Koch (1990, 1995) through their quite different methodology for investigating the phenomenology of vision.[8]

[8] In Jackendoff 1987, I made an explicit claim about the functional correlate of visual consciousness, identifying it with an enriched version of Marr's (1982) $2\frac{1}{2}$D sketch. Since Marr's death, functional/computational accounts of vision have fallen on hard times, and visual theorists have in particular disavowed Marr's levels of visual structure. For this reason, I have to be vaguer now about vision than I was 20 years ago. However, Prinz (forthcoming) offers a contemporary version of the Intermediate-Level Hypothesis for vision (much more specific than that of Koch (2004), who also endorses the idea of an intermediate-level theory), along with a spirited defense against numerous possible objections. He also speculatively extends his account to other modalities.

Thus overall we might think of the functional correlates of the conscious field as forming a sort of “horizontal layer” of cognitive structures in the mind, cutting across “vertical” modalities; this layer is approximately at the level of “perception” rather than either “sensation,” which is too shallow, or “cognition,” which is too deep.

Even if this is correct, it still seems rather senseless. Why should consciousness be associated with the “horizontal layer” of *perception*, rather than with cognition or thought? Many people seem ready to reject this position because it seems to accord consciousness no function. But I think it’s sometimes a mistake to demand a story about the *function* of something before we know what it *is*. In fact, I think this has been a major mistake in the case of consciousness.

To be a little more positive, let me offer some speculation. Let’s return for a moment to the idea that consciousness has something to do with executive function—that it steps in when processing becomes difficult (see citations in section 3.2). Where does this idea come from? A frequently cited intuition is that when you are learning a complex action—say driving a car or playing the oboe—a lot of the mechanics are present in awareness. But once you learn the action well, it becomes automatized, and it fades from awareness, except when you run into an emergency situation. This leads to the intuition that automatized actions are unconscious and that consciousness has the function of dealing with unusual or difficult situations.

However, let’s contrast this case with one that I have *never* seen addressed in the literature. Imagine you’re relaxing on the beach, without a care in the world. There are no cognitive difficulties here, no unusual conflicts to be solved, you’re just enjoying the sun and the waves and the people passing by. Now notice: there is certainly no lack of consciousness here (unless you fall asleep)—there are plenty of very pleasant qualia. Thus in this case we can’t responsibly claim that the purpose of consciousness is to solve problems.

I would like to suggest a different interpretation of these intuitions, one that takes into account the close relation between consciousness and *attention* (as stressed by Posner (1994), Mack and Rock (1998), Dehaene and Naccache (2001), Driver and Vuilleumier (2001), Kanwisher (2001), Hardcastle (2003), Lamme (2003), Koch (2004), and Dehaene et al. (2006), for instance). The central observation is that our attention is attracted and held precisely on things that we are conscious of. If we’re paying attention to something, it’s something we’re conscious of; and conversely, if we're not conscious of something, we can't be paying attention to it.[9]

In order to explore this connection, we have to make a distinction between two senses of "unconscious." On one hand, there are phenomena of which you are deeply unconscious, for instance the condition of your semicircular canals. You may experience dizziness as a result of what's going on in the semicircular canals, but the two are not the same thing. Similarly, you are totally unconscious of the saccadic movements of your eyes. You may experience shifts of attention, but again the two are not the same thing, and it is amazing to see a record of your exact eye movements. Moreover, you simply cannot pay attention to your saccades or your serotonin uptake. I'll call such phenomena *in principle unconscious*. As stressed in section 2.2, one's f-knowledge of language also belongs in this department of the unconscious mind.

On the other hand, consider the situation when you're driving on automatic pilot: you are paying attention to your conversation or what's on the radio or the scenery or your musing about the meeting coming up. But are the mechanics of driving *totally* unconscious? My sense is that they may or may not be. Even if they are at the moment unconscious, you can draw attention to them deliberately, or you can have your attention drawn to them by something in the environment, say the feeling of the car skidding. And often, even while you're conversing, they may be so to speak around the fringes of your awareness. When you are lying on the beach, your attention is unfocused and undirected. It is attracted by whatever happens by: the birds overhead, the smell of food, the sound of the waves and of kids playing, the people walking by, the memory of being at the beach many years ago. And when you confront a traffic emergency while driving, your attention is narrowly focused, directed on what is happening and what you have to do next—to the extent that you may not even notice something important in plain sight (such "inattentional blindness" is documented experimentally by Mack and Rock (1998) and Driver et al. (2001)).

[9] Larry Weiskrantz has brought my attention (!) to some phenomena in blindsight where subjects apparently attend to parts of the blind field and thereby enhance their performance (Kentridge, Heywood, and Weiskrantz 1999). A possible reply is that one can attend to a *location* independently of whether it contains any perceived entities (e.g. "Keep your eyes on that part of the sky"), thereby increasing the sensitivity of visual perception at that point. This case would then contrast with neglect, where attention cannot be directed to the blind field (Ramachandran 1995; Kinsbourne 1998).

These observations suggest that there is a gradient in awareness. On one end there is the indisputable field of *central* (or *focal*) *awareness*. On the other end is what is going on in automated driving: it consists of phenomena that are attendable but currently unattended. I'd like to call this *potential awareness* (you can call it anything you like, as there seems to be considerable debate about whether we should call these cases conscious or not; Dehaene et al. (2006) call it "preconscious"). In between potential awareness and central awareness is a penumbra of phenomena that we are "vaguely aware of," which we might call *fringe awareness*. Central awareness is reportable, but fringe awareness, although it is "there" for us, is not necessarily reportable. A lot of the lying-on-the-beach awareness is in the fringe.

My use of the term "fringe awareness" here seems to correspond to James's (1890) sense of the term. Koch (2004) calls it "gist perception." Block's (1995) "access consciousness" corresponds closely to my notion of central awareness; his "phenomenal consciousness" seems to include fringe awareness and potential awareness together. My impression is that the literature does not on the whole distinguish potential from fringe awareness. Some authors are thus forced into claiming that one is aware of phenomena in potential awareness but forgets them too quickly to be able to report them. Such an account falls afoul of Dennett's (1991) argument about "Orwellian" versus "Stalinist" theories of unconscious content: it is only twisting words to say that something is conscious but you're not aware of it. The problem is solved by admitting that awareness is in part a matter of degree, including fringe as well as central awareness—but not including potential awareness.

The phenomena in central, potential, and fringe awareness are not distinguished by having different sorts of mental structure. Consider a conversation going on at the next table in a restaurant: it is just background noise in fringe awareness—until you start eavesdropping, at which point it achieves central awareness. The mental structures are all of the same character, namely phonological structure. Moreover, the unattended structures must be processed deeply enough that your name or some other key word (say *cognitive science* or *Chomsky*) can attract attention. Iwasaki (1993) discusses the processing going on in fringe awareness; Driver et al. (2001) and Cavanagh, Labianca, and Thornton (2001) show that visual processing goes on even in parts of the visual field subject to inattentional blindness, to the extent that the structures in potential awareness are segmented and parsed by Gestalt grouping (i.e. perceptual parsing).

So how does attention make the difference between central, fringe, and potential awareness? The literature on attention suggests an answer in two parts. First, attention allocates processing resources in the brain, claiming more resources for phenomena of some urgency and taking resources away from phenomena of less current importance. This might be realized in the brain in terms of greater blood flow in the relevant regions, more neurons being activated in the relevant regions, lower thresholds of activation (Koch (2004) speaks of "turning up the gain"), greater amplitude of activation (ffytche 2002), or greater resonance among brain areas (Lamme 2003; Dehaene et al. 2006). The functional consequence of having more processing resources is processing that is finer-grained and more refined (the "attentional amplification" of Posner 1994). In turn, since this processing is developing the content features that support the qualia in consciousness, awareness of attended entities is more vivid and immediate. The consequence of having fewer processing resources would be the opposite: less detailed processing, hence less vivid awareness.

This seems to me to correspond precisely to the phenomenology: the more tightly your attention is focused, the more vivid the attended part of the field seems and the more attenuated are the fringes of awareness. And in situations where your attention is not particularly focused, such as lying on the beach, your awareness is to some degree more uniform across the conscious field. In other words, the difference between a phenomenon in central awareness and one in fringe awareness is the degree of attention actually directed to it at the moment.

This line of reasoning accounts for some of the standard intuitions about consciousness, but in a different way. The capacity limitations of working memory (the "global workspace") require processing resources to be allocated to some restricted portion of one's perception and activity. The executive function of allocating resources is performed not by *consciousness* but by *attention*. At the same time, the differences among attended activities, automated activities, and lying on the beach follow from the way attention is directed in each of these circumstances. On this story, then, the issue of executive intelligence is more about attention than about consciousness.

But there is a second function of attention, stressed by Pylyshyn (2000), Cavanagh, Labianca, and Thornton (2001), and Cavanagh and Alvarez (2005), among others. It turns out that attention is closely bound up with the assignment of "indices" to percepts. A percept's index is what enables it to be tracked over time as it changes position and even properties; it is the index that makes a percept count as "the same thing" with a history over time. (Sometimes the index is called an "object file.") One of the limitations of visual working memory is that it can track only about four indices at a time (the number has been reduced since Miller's (1956) estimate of 7 ± 2). Opinions differ at the moment about whether indices are assigned by attention or whether the assignment of an index is what permits a percept to be attended, but either option will do for present purposes.

What is the function of perceptual indices? An index is what gives a percept its "that-ness"—it is not just a collection of perceptual features, but an *individual*. Computationally, an index is what enables percepts in different modalities to be bound together, so that in viewing a movie, for instance, one can hear the noise coming out of the loudspeaker as the voice of a character on the screen. And, crucially, the assignment of an index is essential for linguistic reference, as stressed in Pylyshyn 2000 and Jackendoff 2002a, chap. 10. Hence the *reportability* of a percept depends on the assignment of an index, and therefore on attention. In turn, reportability is one of the crucial characteristics of central awareness; we can now attribute it to the functioning of attention. Phenomena only in potential awareness, such as occur in inattentional blindness and backward masking, never achieve an index; so, although they can affect behavior, they cannot be reported.

Let me follow this line a bit further. We intuitively think of attention as being attracted by and directed to entities in the environment. But if we ask how this mechanism actually works causally, we realize immediately that attention cannot possibly have direct unmediated connection to the environment. Rather, it has to be driven by processes going on in the brain. That is, the actual entities in the environment cannot direct and hold attention: only some kind of mental structure stimulated by entities in the environment can do this.[10]

So which kinds of mental representations are the anchors for attention? I would like to suggest that the relevant representations are precisely the perceptual "layer" that constitutes the functional correlate of consciousness. Why this perceptual level of representation rather than sensation or cognition? That's harder to answer. I suspect this is the level in perception

[10] Pylyshyn (2000) slips up here, when he says (p. 200), "One possible solution [to tracking objects] is to have a pointer from a representation of an object to an actual object (in the scene), which would act as a demonstrative reference." The brain simply cannot point to an actual object in the scene. Everything has to be done internally; see section 1.2.

where items in long-term memory can first be brought to bear in trying to make sense of perceptual input. In language, for example, the processing of the acoustic signal into putative phonemes can be done pretty much bottom-up. But having reached the phonological level, it is now necessary to call upon the lexicon in long-term memory, to see how the string of sounds is broken up into words and what those words would mean. That is, this is the point where top-down knowledge begins to play a role (see section 2.10.2 and Jackendoff 2002a, chap. 7). My guess is that visual perception has a similar breakpoint. It is only once the visual field has been provisionally segmented into contours and surfaces—the level that I take to be appropriate for visual qualia—that it is possible to call upon long-term memory (the "visual vocabulary") for knowledge about what sorts of objects these contours and surfaces could be the contours and surfaces *of*.[11]

So perhaps we have a story about the importance of the levels of mental structure that constitute the functional correlates of consciousness: they are the first levels in perception where top-down processing comes into play, and they are the levels of structure where attention is attracted and anchored. This does not tell us why consciousness per se is localized there—why these structures rather than others are the source of qualia—but it is a start.

3.5 How Language Enhances Thought

Now let's turn back specifically to the interaction between language and consciousness. I've tried to convince you that the form of thought (here called "conceptual structure") is unavailable to consciousness—it is *in principle* unconscious, to use the term of the previous section. Unlike, say, the phonetic form of words, no amount of attention or introspection is going to yield the computational form of thought. Our thoughts are revealed to us primarily through linguistic imagery, which is correlated with phonological structure. Of course, our visual and proprioceptive im-

[11] This does not preclude feedback from this level of processing to more "sensory" levels such as V1. It is just that there are no V1-like structures stored in long-term memory to be directly invoked in the course of processing.

Prinz (forthcoming) offers a different (and complementary) possible reason for attention to be yoked to the intermediate level in vision: this is the first level in processing that gives any key to what is "out there," and, because it is egocentric rather than allocentric, it is more quickly mapped to appropriate rapid body action.

agery also give us evidence of our thoughts, but my impression is that the average academic is overwhelmingly dominated by linguistic imagery, the unceasing voice in the head. (It might well be different for carpenters and painters and musicians, whose thinking takes place predominantly in domains that are better imaged nonlinguistically.)

What about the consciousness of other primates? As section 3.2 suggested, other primates have thoughts: conceptual structures whose content concerns matters like getting around in the physical environment and engaging in complex social interaction. On my story, the apes' conscious manifestation of these thoughts will be only in the form of visual, auditory, and proprioceptive imagery, just like we have when we imagine seeing and hearing things and performing actions. (On the other hand, humans are not necessarily the only ones with an extra modality of awareness. Bats and dolphins likely have percepts and imagery arising from echolocation; and dogs' olfactory awareness is doubtless far richer than ours.)

But now notice what our specifically human conscious modality of linguistic imagery does for us. Under the analysis of the previous section, imaged language in the head gives us something new to pay attention to, something unavailable to the apes—a new kind of index to track. And by paying attention to imaged language, we gain the usual benefit: increased power and resolution of processing. This increase in power extends not only to the phonological level, but to everything to which the phonology is bound, in particular the meaning. As a result,

Hypothesis 5 Being able to attend to phonological structure enhances the power of thought.

Here are five ways in which attention to linguistic imagery permits discriminations that are impossible with any other form of imagery (some of these were alluded to in section 3.2):

- All sorts of imagery permit one to attend to a token entity in the environment, but only language permits one to attend to *types* as well as tokens. An entity in the environment can be denoted either by an expression like *Freddie*, which regards it as a token individual, or by an expression like *that dog*, which regards it as a member of a category. Words can be used thereby to help pick out conceptual categories: word constancy indicates kind constancy, which aids inference.
- No other sort of imagery permits a token and a type to be explicitly related by *predication*, as in *Freddie is a dog*. Of course, animals make categorization judgments all the time—but they cannot attend to the act of

categorization per se, as a predicative sentence makes possible. Moreover, predication can be used to display relations among types, as in *A poodle is a dog*. Thus only with language is it possible to attend to relations among categories.

- No other sort of imagery permits one to explicitly attend to *lack* of information about the world, as expressed by questions: *Is there a doctor in the house?* Nor can any other sort of imagery permit attention to what is *not* the case, as in *Freddie is not an elephant*. Nor can any other sort of imagery permit one to attend to modality: *It might rain*; *Suppose it rains*. Nor can any other sort of imagery permit attention to specific times other than the present, as in *Freddie was happier last week*.
- No other sort of imagery permits one to attend to the connections among situations. Consider a construction like *Joe is taking an umbrella because it's raining*, which explicitly encodes a relation between two propositions. Each of the propositions can be encoded in terms of a visual image—you can visually imagine both the rain and Joe's taking an umbrella. But, although there may be a relational valuation, a feeling of connectedness between the two, the connection between the two propositions definitely cannot be imaged visually. The word *because* makes the connection *consciously* explicit and isolable, and therefore allows you to pay attention to this connection, question it, look for evidence for it, and so on. A simple conditional sentence like *If it's raining, you should take an umbrella* expresses a connection between two different modal situations and thereby transcends visual imagery in three respects at once.
- No other sort of imagery permits one to attend to the valuation features of a percept—to compare and distinguish percepts from images and illusions, to consider familiarity or novelty, to express affective or emotional attributes of a percept (this is different from expressing one's own emotional states!), or to consider beliefs, intentions, and desires. (The conceptual counterparts of some valuations will be discussed in chapters 6 through 9.)

In other words, there are three factors that together make it possible for language to enhance thought. First, language is far more explicit in encoding the combinatorial structure of thoughts than any other modality of experience or expression. Second, linguistic encoding includes phonological structure, which is a functional correlate of consciousness and therefore is available as a locus of attention. Third, attention to a linguistic utterance or a linguistic image makes it possible to process the corresponding thought in more detail and with more precision or resolution. Thus, although we are not *directly* conscious of thought, language allows us to be *indirectly* conscious of thought in a way that adds power and precision to thought itself.

An unexpected confirmation of this conclusion comes from sign language. During the 1980s, the revolutionary government of Nicaragua created educational institutions for the deaf, and as a result, numerous deaf individuals who had had no previous exposure to language were suddenly thrown together as a community. The astounding consequence was the emergence in this community of a new signed language (now called "Idioma de Señas de Nicaragua" or INS), and since then the language has grown rapidly in richness and complexity (Kegl, Senghal, and Coppola 1999). Among the speakers of this new language are individuals whose first exposure to language came relatively late in life. My hypothesis predicts that before the acquisition of language, they should have been unable to experience their thought, at least in the same richness that others enjoy. And indeed, in a BBC documentary, one speaker says (in the English translation), "I didn't even know what it meant to think. Thinking meant nothing to me." Of course, he had to be able to think before he learned to speak; after all, he wasn't a Cartesian automaton. But he just wasn't aware of it.

3.6 Concluding Remarks, including Evolution of Language Again

All right. After this long and complex journey, we have arrived at the point that everyone has always intuitively wanted to make (all the way back to Descartes at least): that by virtue of having language, humans can think. Well, not quite: on my story, by virtue of having language, humans can think *much better*—and they can be much more aware of their thinking. My position differs from the mainstream intuition in claiming that we are aware of our thought only indirectly, via the phonological structure that language associates with thought. Whether you like this conclusion or not, I want you to notice how the detailed analysis of linguistic structure plays an important role in the argument. It shows we can't simply regard "language" as an undifferentiated whole.

Crucially, the articulation of linguistic structure into phonological, syntactic, and conceptual structure allows us to ask about the functional correlates of consciousness: what aspects of language are most responsible for the qualia that make up the experience of language. It is this question, plus serious attention to the structure of the phenomenology, that leads to the contrast between the present theory of consciousness and many others in the literature. In particular, consciousness cannot be identified with thought (that's conceptual structure), with the contents of a global workspace (that's working memory), with executive function (that's more like attention), or with special properties of neurons (that's the whole nervous system). Consciousness must be distributed among the areas of the brain responsible for the level of perception in all modalities, plus the cross-modal areas responsible for valuation. It remains to be seen if the details of the linguistic case can be carried through in other modalities (see Prinz, forthcoming, for an attempt, particularly in vision).

The present approach also permits a new tack on the relation between evolution of language and evolution of human thought. Section 2.10.4 raised the question of what adaptive function led to the emergence of the language faculty, and endorsed the commonly held view (argued in Pinker and Bloom 1990) that language emerged in the service of enhancing communication among individuals. However, another widely held view is that language emerged in the service of enhancing thought, perhaps through inner speech. Pinker and Jackendoff (2005) argue against this position, on the grounds that (a) there is no adaptive reason for thought qua thought to be cast in a form that is so strongly constrained by the demands of motor performance, in particular its linearization and its tuning to the properties of the vocal tract, and (b) there is no way for inner speech to take place without a vocabulary and grammar that is learned through communicative interaction with others (as observed by the Nicaraguan signer quoted above).

The hypothesis of linguistic consciousness developed here offers a further line of confirmation. The moment language emerged as a communication system, it necessarily had to involve the level of perception—perceiving both one's own and others' overt speech—and thus it automatically afforded a new locus for attention. Moreover, like all other kinds of perception, language perception automatically had an imagistic counterpart, namely verbal imagery or inner speech. In turn, by virtue of the architecture of the mind, inner speech and its capability for enhancing thought would have been automatic consequences of the emergence of language as a communication system. In contrast, the reverse would not have been the case: enhancement of thought would not automatically lead to a communication system. In other words, if anything was a "spandrel" here, it was the enhancement of thought, built on the pillars of an overt communication system—it was not the communication system itself. Of course, this does not preclude subsequent steps of coevolution, where the adaptivity of enhanced thought played a role in shaping further evolution of the language faculty.

This story basically sharpens the common intuition that there is an intimate relationship among the development of language, the development of genuinely human thought, and the development of civilization. It makes it possible to more clearly pose questions about this relation: What sorts of concepts are necessary in order to achieve civilization? Which of them could have been prelinguistic, and which require language in order to be formulated and transmitted? Exactly how does the awareness and attention afforded by the phonological modality help support richer inference? And what evidence can we find in the paleontological record for the emergence of such inference in our species? Barring the unlikely possibility of building time machines, we should not expect much beyond informed speculation. But given the intrinsic fascination of these questions, people are not going to stop asking them. The approach suggested here at least offers the hope of making the speculation somewhat more informed.

Chapter 4

Shaking Hands and Making Coffee: The Structure of Complex Actions

4.1 Introduction

Cognition is not an end in itself. The reason for having a brain is to be able to act. Cognition only does the organism good if its results can be put to use in formulating courses of action. And there is more to action than motor control, especially in the organization of complex (multistep) actions. This chapter examines two banal complex actions that we take absolutely for granted: shaking hands and making coffee. I will use these examples as a vehicle to explore some of the complexity in the systems of knowledge and processing that underlie action.

I undertake this exploration with two ulterior motives. The first is that much of the rest of this book deals with intention, social interaction, and norms governing social interaction such as morality, obligations, and rights. Since all of these involve the determination of action, it behooves us to get a handle on what actions are like and how they are structured.

My second ulterior motive is to use the structure of action to help address the question of how much of the human language faculty is a cognitive specialization and how much has been "borrowed" from other faculties, either in the contemporary brain or in the course of evolutionary development from primate ancestors (Hauser, Chomsky, and Fitch 2002; Fitch, Hauser, and Chomsky 2005; Jackendoff and Pinker 2005; Pinker and Jackendoff 2005). Just looking at brains or at behavior is not enough: in order to compare another capacity with language, we need something comparable to the detailed account of linguistic structure sketched in chapter 1. Such an account is available for music (Lerdahl and Jackendoff 1983; Jackendoff and Lerdahl 2006). But music, like language, is a very special human capacity. The capacity for complex action presents a more ecologically robust candidate for comparison. It is a temporally and hierarchically structured domain, fundamental to human life, one that is undoubtedly shared to some degree with other species.

My methodology here will be to muster a lot of commonsense observations and to attempt to give them some theoretical structure, showing where the inquiry intersects with problems that have been addressed in the literature on the performance and perception of action and in the literature on robotics. Clearly, the story is very preliminary. To flesh it out, one would want to work out many more such examples and, more importantly, work out research methodologies for examining the hypotheses experimentally.

The leading questions of this enterprise might be framed as follows:

- What is the repertoire of structures that can be composed to form complex actions? This question is parallel to the question in linguistic theory of the *grammar* used to compose sentences.
- In any particular action, what are the stored parts out of which it is built? This corresponds to the issue of the *lexicon* in linguistic theory (the artificial intelligence (AI) literature has used the terms "library" (Fikes and Nilsson 1971; Sacerdoti 1977) and "Actionary" (Badler et al. 2000)).

These two questions concern the *form* of the cognitive structures underlying complex actions, in a sense corresponding to the inquiry into linguistic competence. These structures play a role in the following four questions as well, which correspond to questions about how the grammar is put to use in linguistic performance:

- How are complex actions constructed online in preparation for execution? This is the problem of *planning*.
- How are complex actions executed online? This question and the previous one together correspond to the issue of *language production*.
- How are complex actions by others recognized? This corresponds to the issue of *language perception and recognition*.
- How are complex actions specified by linguistic descriptions? For instance, how does one formulate instructions to tell someone else how to perform a complex action?

Furthermore, as in linguistics, one can ask the following learning-theoretic questions:

- How are new stored complex actions acquired? How are the means of composing complex actions online acquired?
- Is there an innate basis that provides the overall form in terms of which complex actions are built and learned?

I won't deal with these questions systematically, but they will be constantly addressed in different ways throughout the chapter. Another question, which I will *not* address, is how all this is neurally instantiated, not just the motor program but all the cognitive apparatus that leads up to it (see Humphries, Forde, and Riddoch 2001 for discussion of "action disorganization syndrome," in which complex actions are disrupted by brain damage).

4.2 Shaking Hands

The simple act of shaking hands reveals several salient features of actions. First, it has structure in two separate cognitive domains, the physical and the social, a more general issue we'll take up in chapter 5. Second, the structure in each of these domains and the interaction between them exhibits interesting complexity. Third, shaking hands is a cooperative action, requiring coordination between actors.

Shaking hands is of course a social convention. Young children don't seem to "get it"; it's something they have to learn to do. But just to say "Shaking hands is a social convention" doesn't explain anything. At bottom, we must store something in memory that might be characterized as our "knowledge of shaking hands." This has to have at least two parts: *how* to do it and *when* to do it. Both of these have to be learned, of course. In addition, looking ahead to chapter 5, there is the more basic question of what kind of knowledge a social convention is.

One's stored knowledge of shaking hands surely does not provide a full set of instructions for all possible situations, which may vary considerably in both physical and social dimensions. So the knowledge must be stored in some schematic form, the "shaking-hands action *type*" or "*schema*," which is modulated or adapted or adjusted to suit particular token circumstances. In this respect, action categorization is much like perceptual categorization: not all dogs or tables or bicycles look exactly alike, and one's stored encoding of the category must be modulated or adapted or adjusted to fit particular tokens one may encounter. No surprises here.

On the other hand, since an actor ultimately has to use the stored schema to produce a token action, the theory of action has to include principles of composition that allow the actor to modulate, adapt, or adjust the schema to suit the particular circumstances and to realize it as a complex of motor instructions, presumably in working memory. Speech production is a familiar case: the motor production of a stored word is modulated by the intonation contour, the overall speech rate, the emotional tone, and possibly the fact that the speaker is chewing gum or smoking a cigarette.

4.2.1 The Social Plane of Shaking Hands

As chapter 5 will document, humans conceptualize people, actions, and artifacts in terms of two independent but interacting planes (or *tiers* in the sense of chapter 1), the physical and the social. The physical plane involves concepts of physical objects moving in space and exerting forces on each other. The social plane involves concepts of *persons*: individuals with whom we can have social relations. People and their social actions are typically encoded on both planes simultaneously.

One's knowledge of shaking hands of course involves the physical plane. One has to know how to actually perform the action: extend one's hand appropriately, grasp the other's hand, shake, and let go. I'll discuss this in a moment. But the *reason* for performing this action is rooted in its social significance: it is an assertion or confirmation of social connection or solidarity.

There are at least five circumstances where one may shake hands: greeting, taking leave, closing a deal, introducing oneself, and congratulating someone.[1] The first four are symmetrical, in that either participant can initiate the action. Congratulation, however, requires that the person doing the congratulating initiate the action, not the person being congratulated.

In order to know when it is appropriate to assert solidarity or social connection, the actors must be able to identify the larger social frame in which an action of greeting, taking leave, and so on can be identified. You don't understand shaking hands if you are just constantly shaking hands inappropriately or don't know to do it when it's called for. More generally, we might ask why we greet and congratulate each other, anyway. Why is a physical gesture of social connection so necessary as part of such actions? And what is the nature of the "social connection" that shaking hands symbolizes? I will begin to address these questions in chapter 5.

The same social significance can be attached to different physical gestures. In some cultures, mutual bowing takes the place of shaking hands;

[1] Notice that being introduced sometimes coincides with greeting, but not always. You're casually conversing with a stranger on an airplane, and at some point you somehow mutually decide it's appropriate to tell each other your names. Suddenly you can't resist shaking hands.

in others, high-fiving. In some cultures on some occasions, the appropriate gesture between a man and a woman is the man kissing the woman's hand. In American culture, a certain degree of intimacy permits (or even demands) the substitution of a hug and/or a kiss; the choice depends in part on the gender and sexual orientation of the participants (heterosexual men probably require the greatest intimacy in order to permit this substitution). The European version calls for kisses on both cheeks (or three iterations in Holland). Most of this is so familiar as to be nearly transparent. Nevertheless, we can already see that choosing the right physical action to "express the social meaning" requires complicated knowledge of the status of the participants and their cultural expectations, including least (a) culture, (b) gender, (c) formality of the occasion, (d) degree of intimacy between the individuals.

Modulations of the physical action can have social meaning too. Most blatantly, declining an offered handshake is an obvious snub. So is offering a more formal action when a more intimate action is appropriate. Conversely, offering too intimate an alternative can be seen as social aggression. The intensity of the handshake and the character of the accompanying eye contact, whether consciously produced or not, also carry social meanings such as assertions of dominance or of disengagement.

How much of this is stored as part of the "knowledge of shaking hands"? Some parts might come from elsewhere, simplifying the "shaking hands" schema in memory. First, one must independently judge degree of intimacy for other purposes in social interaction, for example in the choice of formal or informal second person pronouns (*vous* vs. *tu* in French) or honorifics (e.g. in Japanese), as well as overall choice of linguistic register (formal vs. casual). Second, since hugging is primarily a sign of affection, the choice between shaking hands and hugging depends in part on whether overt displays of affection are independently judged appropriate. For instance, hugging is less appropriate in the context of a formal ceremony than when welcoming a visitor into one's home. Third, the modulation of the intensity of the handshake and accompanying eye contact probably falls out of more general principles for signaling approach, avoidance, and like and dislike. But if we leave these matters out of the handshaking schema in memory, then we put the burden on the combinatorial system, which must integrate judgments of intimacy, formality, gender, and attitude into a fully fleshed out script for performance.

On the other hand, the need for some of these judgments can be shortcut by storing in memory how one interacts with particular individuals (my brother doesn't like to hug, my sister-in-law greets me with a kiss on the mouth, etc.). I take such particular memories to be analogous to one's memory for particular objects: one has not only a schema for tables in general but also one for the table in one's kitchen (an issue to which we return in section 4.4). A parallel relation between stored schemas and stored special cases appears in linguistics in terms of "inheritance hierarchies" (section 2.8). Similar notions of inheritance are invoked in the AI literature on actions (e.g. Kautz 1990; Pollack 1990; Kipper and Palmer 2000).

4.2.2 The Physical Plane of Shaking Hands

For a first approximation, the physical action of shaking hands can be described in terms of a sequence of five subactions:

(1) reach hand to other person > grasp other's hand > shake >
 release grasp > withdraw hand

Such segmentations for a variety of actions are attested by Tversky, Zacks, and Lee (2004). As we will see in a moment, there is more structure to shaking hands than this sequence. But first let's ask about the theoretical status of (1).

If (1) is one's knowledge of the physical aspect of how to shake hands, it has to be stored in some form in long-term memory. Suppose that at some point it is called up into working memory as a possible action and incorporated into situation-specific context: this has the status of a *plan*.[2] Suppose that the actor commits to the plan; it now gains the status of an *intention*. Finally, for the schema to be carried out, it has to be instantiated in actual motor instructions that are read off in temporal order; it thereby becomes a *voluntary action*. In other words, our intuitive ontology of plans, intentions, and voluntary actions corresponds nicely to the various roles that an action structure can play in cognition and execution. (We return to planning, intending, and voluntary actions in chapter 8.)

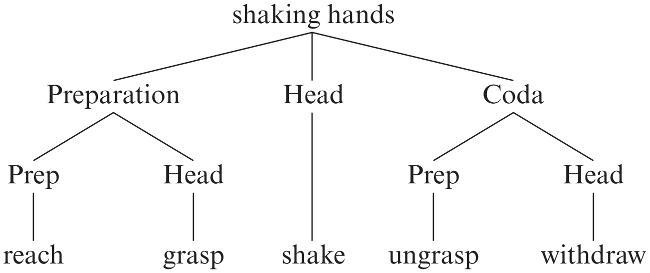
Now consider the sequence in (1) in more detail. It makes sense to structure (1) into constituents, where each constituent has a subaction or a subconstituent as *Head*. The Head of the entire sequence in (1) is the actual shaking of hands: that's the point of the sequence, the subaction that the whole thing is *for*. Reaching and grasping then constitute a

[2] A linguistic fine point noted by Bratman (1990): I mean here a "plan *for* shaking hands"—that is, an action under consideration. *I plan **to** shake hands* expresses a commitment or intention, the next step toward execution.

Preparation for shaking; and releasing and withdrawing constitute the "finishing-off" or *Coda.* This organization is reflected in an interesting asymmetry in the way we describe the subactions. One reaches and grasps *in order to* shake; but one certainly does not shake in order to release and withdraw. Rather, one shakes in order to affirm social connection; and one releases and withdraws in order to return to the original state, to "put things back in order."

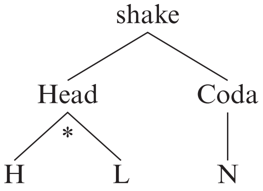
Is there any organization within the Preparation and the Coda? I think so. It seems reasonable to say that one reaches in order to grasp; that is, reaching is preparation for grasping. And it seems reasonable to say that one releases grasp in order to withdraw. The overall structure, then, can be encoded as a tree structure like (2). Trees along these lines appear in the AI literature in Litman and Allen 1990, in the psychological literature in Whiten 2002, and in ethnoscience in Werner and Topper 1976, for instance.

(2)



The "shake" constituent also has more structure. Intuitively (we'd want to check this empirically), there is a sequence of up-and-down motions, oscillations above and below a neutral position, starting by moving upward to high position, and ending by moving upward from low position to the neutral height. This sequence has no fixed length, and it contains no single action that serves as Head. So we might notate this in a tree structure as (3), using H, L, and N for high, low, and neutral position respectively, and using the star as a sign of indefinite repetition (following the custom of the "Kleene star" in formal languages). I'll treat the return to neutral position as a coda of the "shake" constituent.

(3)



Much of this structure requires coordination with the other participant in the course of execution. Your reach is modulated by perceived height and distance of the other participant. You don't begin shaking hands by sticking your hand in the other person's face; you are aiming for a point roughly midway between the two of you—higher if the person is taller, lower if the person is shorter (or seated). The strength of your grasp is calibrated to match the other person's (well, *most* people do this); the tempo and amplitude of the shake must be coordinated; the ungrasp must be pretty close to simultaneous. The modulation of the reach is based on visual perception; from the grasp on, everything is calibrated tactilely and proprioceptively.

The coordinated timing of these actions between individuals cannot be the result of a miracle. Rather, shaking hands is a collaborative action, a *joint action* in the sense of Gilbert 1989, Searle 1995, Clark 1996, and Bratman 1999, and to be discussed in chapters 5 and 8. It is not just me taking your hand and shaking it, nor is it me taking your hand plus you taking my hand. It is us doing this together. Joint action need not involve social intent, for instance as in lifting a heavy object together. But it does create a sense of social connection, through "you and me acting as one."

Clark (1996) points out that a joint action usually requires an "offer" by one participant and an "uptake" by the other. In particular, in shaking hands, one participant must initiate the action and the other must pick it up, very quickly (I'd guess 500 milliseconds or less) and often not consciously. I think that the only time I'm conscious of the cues is when they misfire, for example if I loosen my grip and the other person keeps shaking. The social context likely primes the action, so one's reaction to someone else initiating a handshake is probably faster in an appropriate context than in an inappropriate one. (When I unexpectedly demonstrate shaking hands in the middle of a class, my victim is usually considerably slower to respond.) And there is probably a conventional approximate default length for the shaking stage, which primes the appropriate point to let go. In other words, practically every phase of shaking hands requires initiation and response.

These actions may be concurrent with others. Suppose you happen to see an acquaintance walking toward you on the street. The scenario might go like this: You might extend your hand in advance, while still walking, before you've actually converged to within shaking distance. As you shake hands, you are also talking. Without releasing your grasp, you may turn around each other, so as to continue in the direction you've been going, and then turn away from each other concurrently with ungrasping and withdrawing your hands. And the withdrawal may merge fluidly into waving to each other. So the total movement script in this case superimposes the handshaking schema on a number of other motion patterns, with coarticulation. Like talking while chewing gum, this requires a theory of action composition.

What aspects of the physical pattern are stored in memory as part of the handshaking schema? It's plausible that it's just the "shake" constituent (3). This presupposes that the hands are grasping each other at an intermediate position between the two participants. In order to execute the "shake" constituent, each participant would have to get to the proper position, which would involve constructing an action plan for the Preparation constituent in (2). And in order not to be stuck together holding hands, each participant would have to construct an action plan for the Coda constituent in (2). In other words, the "shake" constituent might be provided with just enough structure on its own for the rest to be constructed online. Again, for this to be feasible, it is necessary to provide the theory of action with the potential for action composition, that is, putting together a complex action from stored parts.[3]

Let's be a little clearer about what constitutes "enough structure." The "shake" constituent cannot now just say, "Move your hand up and down some number of times." Rather, it has to say something like this:

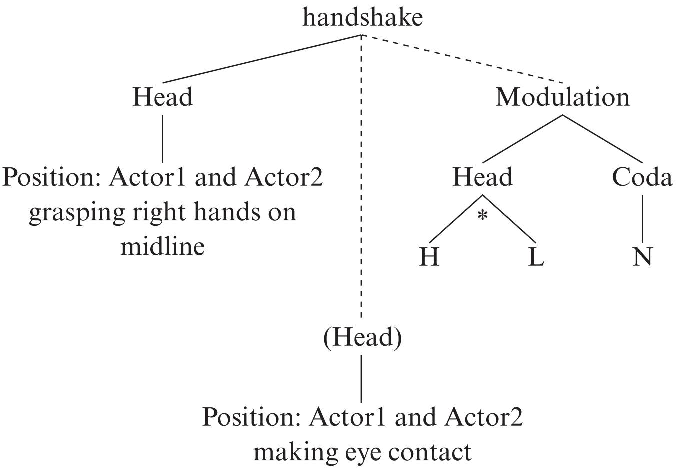
(4) Grasping the other participant's right hand with your right hand in a mutually comfortable midline position,
Move your right hand up and down some number of times.

All the information in the first clause was previously encoded in the endpoint of the "reach" and "grasp" constituents; it is now encoded as the "neutral" position in the "shake" constituent (which required it anyway).

How might a tree incorporate all this information? We want to say there is a basic position that is modulated by up-and-down movements. (5) is an attempt.

[3] The necessity of adapting a schema to current action obviously has a counterpart in robotics. One approach (Badler et al. 2000; Bindiganavale et al. 2000; Kipper and Palmer 2000) encodes schemas in memory as *parameterized action representations*. These are filled in to suit current circumstances with parameters such as objects to be manipulated, manner and duration of action, trajectory of motion, and necessary preparatory actions. Another (probably complementary) approach (Levison and Badler 1994; Shapiro and Ismail 2003) interpolates a level of motor planning representation between action representations and execution. A linguistic analogue might be the level of detailed motor activations as opposed to phonological structure; the latter specifies idealized endpoints of vocal tract movements rather than how to achieve them.

(5)


handshake
Head
Position: Actor1 and Actor2 grasping right hands on midline
Modulation
Head
*
H
L
Coda
N
(Head)
Position: Actor1 and Actor2 making eye contact

The Head is the neutral position; this is modulated by the up-and-down motion, which is notated as a constituent called *Modulation*. The Modulation is connected to the larger action by a dotted line, which is meant to indicate that it is concurrent with the Head rather than successive to it. (If we had three dimensions at our disposal, we could notate this constituent branching off at right angles to the page.) Such a tree configuration is also useful for characterizing motions like waving ("Hold your hand up in the air and wiggle it back and forth"); it also bears some similarity to the sorts of structures posited for hand position and movement in sign language (see various models in Fischer and Siple 1990).

(5) also notates the possibility of eye contact as an additional concurrent action, connected to the complex again with a dotted line. Eye contact is a concurrent action, a sort of co-Head, rather than a Modulation, because it can take place independently of hand grasping. By contrast, the Modulation of moving grasped hands up and down cannot take place without hand grasping.

To sum up at this point: the hypothesis is that a structure on the order of (5) is stored in memory as the physical plane of shaking hands, and that the Preparation and Coda of (2) are constructed online in working memory to provide appropriate transitions into and out of the basic position.

The English description of positions in (5) is of course proxy for an encoding of body position in an appropriate spatial/proprioceptive format. One possibility is the level of spatial structure proposed in Jackendoff 1987, 2002a, itself based to some extent on Marr's (1982) 3D model level, in particular incorporating Marr and Vaina's (1982) generalization to the encoding of object movement. Whatever the details, it is crucially not a propositional structure that maps directly into language, a point to which we will return.

In turn, (5) is linked in both long-term and working memory with the structure of handshaking in the social plane. In the social plane, the structure is "confirming social connection in the context of greeting, etc." The same social structure is also linked in long-term memory to alternative physical realizations such as hugging, high-fiving, kissing the hand, and so on. It's interesting that all of these have a structure rather like (5): they can be characterized as adoption of a position for a brief period of time, sometimes with modulation, sometimes not. Each of these positions requires strategies of approach and withdrawal, which it might be reasonable to suppose are constructed online. And they require cues for coordination: For example, which comes first—does the lady extend her hand, or does the gentleman start lowering his head and pursing his lips while reaching for her hand? It depends, and the choice may have social significance.

4.2.3 Variable Instantiation, Including by Self

I've slipped something into (5) that makes it more abstract than the previous analyses. The Position constituents speak of *two actors* grasping hands and making eye contact, rather than "self and other actor." Now in any event we need a perceptual schema that allows us to recognize two *other* people shaking hands. And maybe (5) is *that* schema. But notice what it would take to use (5) that way. In any particular situation, the people involved are particular people, not generalized people, and we're not just recognizing two people shaking hands, we're recognizing, say, Roosevelt and de Gaulle shaking hands. In order to achieve such recognition, it is necessary to treat *actor1* and *actor2* as variables, which are instantiated on particular occasions by different individuals. In other words, recognizing others' actions as known actions requires online variable instantiation, of the sort familiar from language processing. Here we connect with the event perception literature (Zacks and Tversky 2001). In particular, Cavanagh, Labianca, and Thornton (2001) suggest that familiar complex motion patterns such as walking are stored as "high-level animation" or "sprites," ready to help identify perceived patterns. Shaking hands might be one of those.

Now, what would happen if by chance one of the variables in (5) were filled by *self* (or *ego*)? The result would be the action of "me and other actor shaking hands." Now there is a basic asymmetry about the world:

I cannot, simply by intending it, move someone else's body, but I *can* move my own. That is, there is an eccentric connection between the conceptual formulation of action and the motor system: the motor system can be engaged to formulate a motor program if and only if the actor in a conceptualized action is *self*.[4] If this is the case, (5) could serve to schematize handshaking at the most general level, applying to both the perception and the production of handshakes. The latter could occur only if one of the actors is *self*.

On the other hand, there is often considerable disparity between one's ability to perceive an action and one's ability to perform it. Lots of people can tell a good dancer or basketball player from a bad one, but can't produce the actions with any competence at all. My uncle Bernie was an avid and sophisticated operagoer, but could barely carry a tune. Presumably, in order to use one of these schemas for performing an action, one must link it with mental structures specialized for proprioception and motor control (speaking as usual is a good example). Practicing an action refines these proprioceptive and motor structures; watching someone else do it does not. The disparity between perception and production also shows up in the many actions that one can recognize that one has never performed. For example, I don't believe I've ever kissed a lady's hand, and I've certainly never been a lady having her hand kissed. But one needs some mental encoding of such actions, so one can recognize them and judge if they are correctly performed in an appropriate context.

The idea of stored action schemas that serve both perception and production is not so surprising if we remember that *words* have a similar status. Recall from chapter 2 that these are stored linkages of mental structures, and of course they are used in both language perception and language production. As with actions, one may have perceptual command of a word—be able to interpret it in context—without necessarily having productive control—being able to use it fluently and appropriately. So again the parallel with language proves useful in understanding what we're doing here. The overall point is that a structure like (5) is on the cusp between action *perception* and action *production*.

Pushing on one step further: in the present formulation, being able to imitate someone else's actions requires (a) formulating an action schema like (5) on the basis of perception, (b) substituting *self* for the requisite

[4] This turns up in chapter 8 as a formal condition on "actional attitudes" such as intending: the Actor of the action must be bound to the intender.

actor, and (c) linking the result with motor control structures. Whiten (2002) discusses imitation in the context of complex action, using trees not unlike those above. Here is perhaps a locus of activity for mirror neurons—neurons that respond both to seeing someone else perform an action and to performing the action oneself (Rizzolatti et al. 1996; Flanagan and Johansson 2003; Nelissen et al. 2005). The going wisdom these days is that other primates are considerably less proficient at imitation than humans (Donald 1998; Hauser 2000). It would be interesting to ask where in the chain of mental structures the difference lies (evidently not in the known mirror neurons, since they were first found in monkeys!).[5]

I'll leave shaking hands here. Let's do something else.

4.3 Making Coffee

Making coffee lacks a social aspect but compensates with other complexities. It is representative of a whole repertoire of "practical knowledge" that involves using objects. I want to draw attention to two aspects of such knowledge. The first is knowing the appropriate way to use artifacts. We have a vast amount of such knowledge: even a five-year-old knows how to use shirts, socks, buttons, zippers, Velcro, beds, doors, chairs, telephones, televisions, light switches, crayons, pencils, paper, spoons, cups, sinks, bathtubs, and on and on. Such knowledge links object structures and action structures in memory (and, following Myung, Blumstein, and Sedivy (2006), even affects lexical access for object names). A second aspect of practical knowledge is the ability to put artifact knowledge to use in elaborate sequences of behavior. Think of the whole sequence of routines you go through on getting up in the morning: taking a shower, getting dressed, combing your hair, feeding the cats, preparing and eating breakfast, reading the paper, clearing up the kitchen, and so on. All this takes a tremendous amount of organization, complexity, and flexibility.

Humphries, Forde, and Riddoch (2001) show how these abilities can be disrupted by brain damage. A patient with "action disorganization syndrome" may omit steps in a multistep task, or incorrectly instantiate the

[5] Notice that on this story, imitation does not require theory of mind—all it requires is observation of external action. The story becomes more complex, though, when we consider imitation modulated by perceived goals or intentions (Gergely et al. 1995; Phillips and Wellman 2006).

arguments of actions. For instance, in making tea, such patients may omit putting the teabag into the pot, pour milk into the teapot instead of into the cup, or stir the tea in the pot instead of the tea in the cup. They have more difficulty carrying out novel tasks (“Pour from the cup into the teapot”) than more stereotypical ones (“Pour from the pot into the cup”). They may perseverate on a task, for instance continuing to cut wrapping paper smaller even after they verbally note that it is already too small for the package it is to wrap. Humphries, Forde, and Riddoch further document a double dissociation: there are some patients who can describe the action that goes with a particular object (say a cup) but cannot carry it out, and there are others who can carry out the appropriate action but cannot describe it (this case falls under “semantic dementia”). They take this to show that verbal working memory and the department of working memory devoted to formulating action are distinct. This concurs with the discussion in the previous section, which argued that action structures are encoded in some format such as spatial structure rather than a format that inputs directly to language.

4.3.1 Basic Structure

On to the analysis of the case at hand. The action of making coffee varies depending on what kind of coffeemaker you use. One’s knowledge might extend to certain kinds of coffeemakers and not others. I don’t have a clue how to use an espresso machine; the old percolator in my parents’ summer home requires a different technique from the automatic filter coffeemaker in my own kitchen. So even in this restricted domain, one’s knowledge of action is highly tool-specific. For the analysis here, I’ll take the coffeemaker in my kitchen as the operative example.

Again beginning at the grossest level, the steps can be described as in (6).

$$(6)\ \left\{\begin{array}{l}\text{put in coffee}\\ \text{put in water}\end{array}\right\} > \text{turn on coffeemaker} > \text{wait until coffee is done}$$

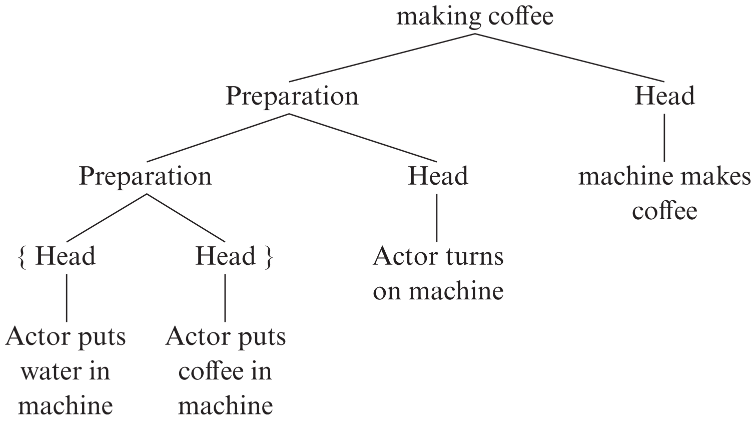
As in shaking hands, there is more structure beyond the temporal sequence. Everything here is preparation for the coffeemaker performing its function while the actor waits. That is, the actor actually doesn’t perform the Head of the action. In order to encode the structure of this action, then, we have to deal with the roles of both the actor and the machine. The actor has to know what the machine is supposed to do (though not necessarily how the machine does it—how many people know how their car or computer works?).

One feature we haven't seen yet is two unordered subactions. Whether you put the coffee or the water in first doesn't matter, as long as you do both before you turn on the coffeemaker. Nevertheless, you probably have a default (or habitual) order in which you do these subactions. I typically do the water first. Similarly at larger scales of action: the order in which I make the coffee, take a shower, and feed the cats in the morning doesn't logically matter, but they are all necessary subtasks of the "getting started in the morning" action, and I have a default order. The situation rather parallels free phrase order in language. I may say *Bill arrives on Thursday at 8* or *Bill arrives at 8 on Thursday*, but on any particular occasion I must choose one or the other.

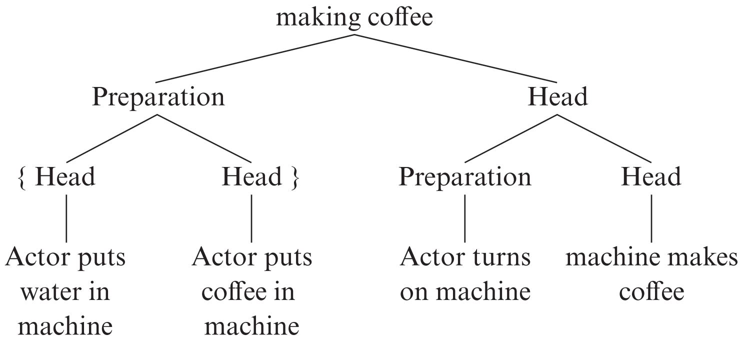
In order to start turning (6) into a tree structure, we have to ask how we want to encode such "habitual but not necessary" temporal order in the grammar of action. Our cases so far have involved necessary temporal order: for example, you can't shake hands till you've grasped hands. These show up in the tree as attachment dependencies: Preparation to Head, or Head to Coda. By contrast, putting in water and putting in coffee are logically independent of each other. I'll notate them as two independent Head branches in the Preparation, surrounded by curly brackets { }.

Another question is whether these two steps are preparation for turning on the machine, or for a larger constituent that consists of turning on the machine and letting it work. (7a) shows the first possibility and (7b) the second. I kind of favor the latter, though without yet having a notion of what would count as evidence. (The evidence would parallel constituency tests in linguistics and might involve either intuitive or experimental procedures.)

(7) a.



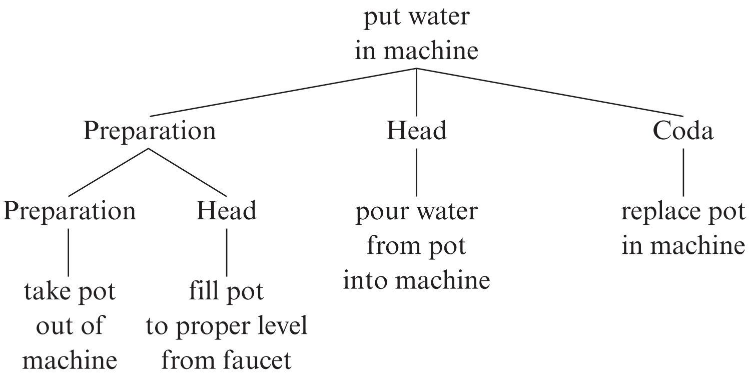
b.



4.3.2 Complexity in the Subactions

Let's look more closely at the subaction "actor puts water in machine." One could perform this action in various ways. The way I do it is to take the pot out of the coffeemaker, put the right amount of water in it from the faucet, pour the water into the back of the coffeemaker, and put the pot back in its proper place. In a tree structure, this might look like (8).

(8)

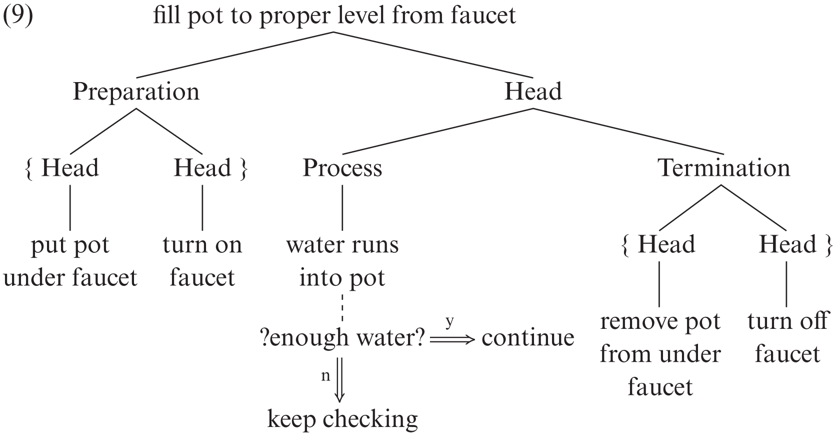


There is more detail, for example in the constituent "fill pot to proper level level from faucet." One problem is what "proper level" is. This is correlated with the amount of coffee one wants to make, which has to be a free parameter in the specification of the task, with perhaps a default setting (for me, six cups). This parameter will show up again in "put coffee in machine," when one has to measure a certain amount of ground coffee depending on the amount of coffee to be made.

An immediate problem arises in how to lay out the tree. The actor puts the pot under the faucet and turns the water on (not necessarily in that order). Then two things happen at once: water runs into the pot, and the actor monitors the water level (which may include some extra steps of checking, such as lifting the pot to eye level). When the level reaches criterion, the actor terminates the process by turning off the water and removing the pot from under the faucet (in some order). Here the process of the water running into the pot seems like the Head, but the actor's removing the pot from under the water doesn't seem like a Coda in the sense in which we've been using the term so far. So let me call these two constituents *Process* and *Termination*, more or less following practice in the linguistic literature on event structure and aspect.

We also need a way to notate the actor's monitoring the rising water level. Our little grammar of actions needs to be supplemented to incorporate *checking steps*. Toward that end, let's introduce the notation *?x?* 'check to see if *x* is the case', where *x* is a state of affairs—here, whether the water level is at criterion. The action *?x?* has two possible continuations, depending on the answer.[6] With all this in place, we might elaborate "fill pot to proper level from faucet" as (9).

(9)



How much of this actually has to be stored as part of the knowledge of making coffee? Filling things to criterion under a faucet is a more general

[6] The resemblance to the "test" step in the old Miller, Galanter, and Pribram (1960) TOTE units ("test-operate-test-exit") does not go unnoticed. Tests are inevitable as a part of any sort of flexible behavior. On the other hand, the particular test in (9) is not a discrete step: it goes on continuously throughout the process until the condition is satisfied.

If a test step is not carried out properly, the result may be perseveration, as in the previously mentioned case of the patient who continues cutting wrapping paper well past the point where it has reached the right size (Humphries, Forde, and Riddoch 2001).

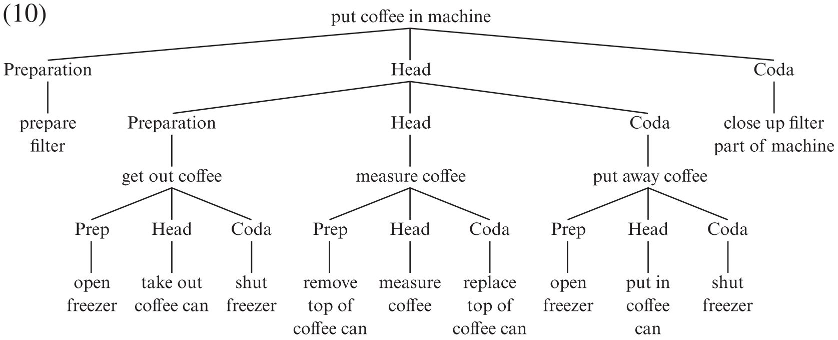
task, and turning the faucet on and off to get water for whatever purpose is a still more general task. So perhaps the knowledge of making coffee only specifies the endstate of the process, namely that the pot has the right amount of water in it, and the rest is constructed online.

Left out of this are even finer details. You have to grasp the pot, disengage it from the machine, and walk with it to the sink; after filling it, you have to walk back to the machine. While the pot is filling, you have to maintain it in a fixed position, which requires applying more force to hold it up as the water adds to its weight. Turning the faucet on and off calls for mechanical knowledge of the faucet: where to apply force, and how much. How the pot disengages from the machine might be part of the knowledge of making coffee with this machine (what else would you use this knowledge for?), but grasping the pot, holding it, and walking surely are not. Nor is your mechanical knowledge of the faucet part of the coffee-making schema; rather, it is its own little schema, full of details of your kitchen faucet and other faucet types you know, perhaps related to a more generic schema that helps you deal with faucets you have never encountered before. All the subactions that are not part of the coffee-making schema need to be recruited online and integrated into the structure of ongoing action.

More interesting is the possibility that the pot needs to be cleaned before being filled. We'll come back to this shortly.

The "put coffee in machine" constituent in (7) offers a few new sorts of details. My coffeemaker has a permanent filter that has to be taken out and cleaned; most people use paper filters instead. In either case, the filter has to be dealt with, the coffee container has to be located (in my case, in the freezer) and opened, coffee has to be measured into the filter, the coffee container has to be closed and replaced in its proper location. The structure, partly expanded, looks about like (10).

(10)



I've left out the cleaning and replacing of the filter (more use of the sink), all of which is subordinate to "prepare filter"; and I haven't touched on the further structure of the lower constituent labeled "measure coffee," where you have to (a) coordinate the amount of coffee with the amount of water and (b) repeatedly scoop coffee out of the can and into the filter, while monitoring the total amount scooped—a more elaborate version of the measuring routine in (9). The scooping, of course, presents more fascinating problems for motor control.

The main thing elaborated in (10) is the sequence of openings and closings. The Codas of putting the top on the coffee can and closing the freezer are not necessary for the task. Rather, as in the Coda for shaking hands, they're necessary for restoring the status quo ante.

Most of the structure in (10) is doubtless not part of the coffee-making routine. Rather, my guess is that the coffee-making routine includes the measuring of coffee into the filter plus knowledge of where you keep the coffee; but all the fetching, opening, and closing is constructed online. Nevertheless, *all these subactions must be performed*, so they must be integrated into the sequence of motor instructions. The upshot is that a performed action, constructed in working memory, has a deeply embedded structure—in this case, (8) and (10) embedded in (7b), and (9) further embedded in (8), plus all the pieces I haven't bothered to expand into their full structure. This tree is complex enough that I will spare us the difficulty of writing it all down in one place.

Three other very general complications need to be fitted in. First, what happens when you're in the middle of making coffee and the phone rings? You have to find a point to break off (I'll bet it's likely to be a constituent boundary), leave a pointer to where you are in the task, answer the phone, then after the phone call return to where you were in making coffee. It's a little like the situation in language when you need to interject a comment in the middle of a discourse. And, just as when your comment gets so long that you lose track of the original thought, you may forget to go back to making coffee after a long phone conversation.

On the other hand, you may interleave the two activities, talking on the phone while continuing to work on the coffee. So we also need the possibility of multitasking, guided by shifts of attention between the two.

Finally, making coffee doesn't require the same actor throughout. My wife may start the coffee and I may finish it up. Or she may measure the water while I measure the coffee. And we both keep track of which steps have been completed. Perhaps this is the result of composing a very general joint-activity routine with the coffee-making routine.

4.4 Building Structure

4.4.1 Parallels to Lexical and Phrasal Structure in Language

In analyzing these actions, a recurring question is what parts of the structure are stored as part of this action per se, and what parts are built online by the instantiation of arguments, by adjunction of other stored actions, and by modulation of the schema to suit the context. A picture is beginning to emerge. The stored schema that pertains specifically to shaking hands encodes the actual shaking action and its social meaning. The stored schema for making coffee is a structure of semiconnected vignettes: measuring a predetermined amount of water from the faucet into the pot, pouring the water from the pot into the coffeemaker, measuring a commensurate amount of coffee from the can, turning the machine on, and letting it do its work. In both cases, a certain amount of "connective tissue" is missing. Where does it come from, and how does it come to be integrated?

We can't take the position that it all comes for free. As people in robotics point out, we can't just leave it all out and expect our robot to perform, using its nonexistent common sense. Often the necessary subactions are recruited from more general processes. One walks from one place to another all the time, not just from the coffeemaker to the sink and back. One reaches for all kinds of things in all kinds of positions, not just for people's hands; one uses the sink for all kinds of things besides making coffee. One opens the freezer for all manner of reasons—and getting things in and out of the freezer is just a special case of getting things in and out of cabinets and cupboards. However, some of *these* actions may have specialized parts, such as how to use your very own kitchen faucet.

Moreover, the construction of actions is not just a matter of sequencing one subaction after another. For instance, in adapting the general process of walking to the goal of getting to the sink, the variable for destination of walking must be instantiated appropriately. For a more complex case, suppose you are going to measure water into the pot, and the pot turns out to contain some dregs of coffee. Then you don't want to go straight into filling it with water; you have to clean the pot first—a more complex Preparation. Similarly, consider what happens if you go to measure coffee and find there isn't any. One possible response is to abandon the task altogether: no coffee for you this morning. But if the need is desperate—suppose you're expecting guests for whom you have to run the "hospitality" routine—you might go out and buy some coffee. In this case, the composed structure of (10) contains, instead of "take coffee can out of freezer," a huge Preparation constituent, which includes driving and navigating and buying routines. And the driving routine may in turn require a search for your car keys, an even more deeply embedded Preparation. So the composed structure of an action can become deeply embedded, in unpredictable ways.

This is beginning to look a lot like the construction of sentences. You can't store in your head all the sentences you can speak: there are just too many of them, in fact indefinitely many. Rather, you store bits of language in memory and combine them in real time to create a sentence that suits your current needs. Moreover, just as you don't usually construct a whole sentence in your head before starting to say it, you probably don't construct a whole complex action in your head before you start performing it. Rather, you add on pieces as you need them (or as you anticipate needing them). Still, in both cases, pieces must be attached in such a way that the output is coherent. In the case of language, coherence amounts to some approximation of grammaticality and meaningfulness, not always achieved. Here we're trying to figure out what the notion of coherence of an action might mean.

Standard views in linguistics have it that the stored pieces are some tens of thousands of *words* plus some unknown but relatively small number of combinatorial *rules* that combine the words into structures. However, according to a view developing independently in a number of different quarters and discussed in section 2.8, there is no strong distinction between words and rules; rather, there is a continuum running from simple words such as *dog*, through idioms such as *kick the bucket*, which have syntactic structure, through regular morphemes such as the English regular plural *-s*, through idiosyncratic constructions such as *the more you read, the less you understand*, all the way to very general principles of combination such as VP ⇒ V – NP. The special cases are related to the more general cases that they instantiate in terms of inheritance hierarchies. Inheritance hierarchies in turn are a more general form of default taxonomic hierarchies, independently necessary for general-purpose categorization: "Since X is a robin, a robin is a type of bird, and a bird is a type of animate, X inherits all properties of robins, birds, and animates unless I have information to the contrary." The basic principle for building sentences from the stored pieces is called "unification" (Shieber 1986; see also section 2.9): the idea is that one clips together stored pieces any way possible, consistent with the constraints each of them imposes, such that every bit of the composed structure is accounted for from one stored piece or another (Shieber 1986).

Such an approach seems apposite for the structure and construction of complex actions from multiple schemas. First consider parallels to the lexicon. As suggested in section 4.3, we store a huge number of action schemas, including among other things how to use all the thousands of artifacts we know how to use. This "action lexicon" is quite possibly comparable in size to our linguistic lexicon. Moreover, as suggested in section 4.2, consider the relation between rather general stored actions, such as how to use a key in a lock, and very specific actions, such as the way you have to jiggle the key in your front door a certain way to get it to turn. This relation is similar to the relation between general linguistic constructions such as PP ⇒ P – NP and idiomatic constructions such as the PPs *in luck* and *in good humor*. One has to add to the general knowledge of keys and locks the extra twist that makes this case special, and this is stored as an idiomatic piece of action knowledge, an additional entry in the action lexicon.

Next consider the stored "semiconnected vignettes" I've posited for the coffee-making schema. There are parallel discontinuous cases in the linguistic lexicon. For instance, the idiom *take... for granted* is certainly a stored unit, but it cannot normally be used in a sentence unless something fills in the gap between its two parts. Another familiar example is the French split negation *ne... pas*. On the other hand, what binds such items together is that their discontinuous parts are connected in the tree structure. This seems also true of the relevant parts of the coffee-making schema.

Turning to the principles of composition for complex actions, we have already talked about building hierarchical structure and instantiating variables. The parallels go further. Among the principles involved in sentence composition are principles of *referential binding*. A particularly apt example in the present context is relative clauses: a relative clause is understood as containing a pronominal element of some sort that corefers with (or is "bound to") the noun that the clause is attached to. For instance, consider the sentence *I drank the coffee that I made last night*. The relative clause is understood as *I made (some) coffee last night*, not, say, *I made a cake last night*. This constraint is reminiscent of situations we have encountered here. For instance, going to the store and buying cheese is an "ill-formed" Preparation for making coffee, and it doesn't make sense to wash a spoon when what needs to be clean is the pot. In other words, the formal device of variable binding that connects a relative clause to its head noun is also about right for connecting objects in subordinate actions to objects in the main actions they prepare. (And if the relative clause gets really long and you lose track, you may do the equivalent of getting to the store and forgetting what you were supposed to buy.)

Another such parallel arises in the necessity of correlating the amount of coffee you measure out with the amount of water you measure out. Such a correlation is directly expressed in a sentence like *For each cup of water, put in one scoop of coffee*—that is, a sentence involving quantification. It's interesting to ask whether one could accomplish such correlations in action structures without having linguistic quantification to support them. I don't have intuitions one way or the other.

To sum up, the grammar of action exhibits significant parallels to the structure of linguistic grammars.

4.4.2 Filling In the Blanks

If actions are stored in memory in highly schematic form, how are all the necessary pieces recruited and integrated into the main schema one is executing?

I'm going to invoke the model of language again. We've established that the linguistic parallel to stored action schemas is stored words and idioms, which can be integrated into sentence structures in an unlimited number of different contexts. What "motivates" their being activated is that the speaker has a piece of meaning in mind. The psycholinguistic evidence (e.g. Dell, Burger; and Svec 1997; Levelt 1999) shows that the "call" to the lexicon is very general, along the lines of "Does anyone in there mean *this*?" We can think of lexical items actively and promiscuously "volunteering" in response to the need for a particular meaning to be expressed. They compete actively with each other for expression; a process of selection (perhaps winner-take-all activation) picks out the one that is actually uttered. If the competition isn't properly resolved, we get speech errors, such as *troblem* for *trouble* plus *problem*.

A similar conception seems appropriate for complex actions. When one decides it's appropriate to shake hands, there is a discrepancy between the position one's hand happens to be in and the position it *needs* to be in (i.e. grasping the other actor's hand). The result is a call to the action lexicon for an action to get one from here to there. The "reach to *x* and grasp *x*" routine stored in memory "volunteers" and is attached as a Preparation to "shake." As part of being attached, its variables are instantiated by the relevant objects in the current situation (the way a verb's variables are instantiated by its subject and object), so that one reaches to the right place and grasps the right thing. Similar mismatches of position are going to get you to walk from the coffeemaker to the sink and back, and to drive to the store to buy more coffee and return (which in turn requires you to walk to your car). That is, the discrepancy "I'm here and I need to be there" can call up a variety of routines depending on the context. This intuitive account is fleshed out in the AI literature on planning (Pollack 1990) and robotics (Badler et al. 2000). For instance, in the latter, a virtual robot (a screen animation) is issued an instruction to walk somewhere at a moment when it is sitting down. In order for the command to be executable, first the robot's action planner automatically appends a Preparation of standing up.

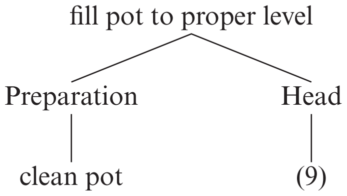
Another very general sort of Preparation based on discrepancy arises in the use of an object that one is not already holding, for instance the subaction "fill pot to proper level from faucet." As Preparation, one must pick up the pot. But in Preparation for *that*, the pot must be located—so I have to know where to find it. My knowledge of my coffeepot is that it sits in the coffeemaker (so I have to know where *that* is), and this was built into structure (8): "take pot out of machine." But this is only part of the coffee-making schema as a default. The coffeepot may not prove to be there, in which case I can resort to some backup knowledge: it might be in the dish drainer or the dishwasher, so I'll look there. Failing that, I have to institute a general search.

The point is that the discrepancy of needing to use a particular object triggers a search for it, and the search can be driven by knowledge very specific to the object. The object belongs *here*, but it might also be in such-and-such other places. Multiply that over all the objects in your house and the knowledge is phenomenally rich. Where do you keep the cake pans? the spare key? Aunt Betty's sterling candle snuffer? And it's not just in your house: Where does your local supermarket stock the pasta? the beer? and so on. Moreover, we also remember episode-specific locations: Where did I put down my reading glasses 10 minutes ago? Where did I leave my car in the parking lot today?

A general routine along the lines of "find object *x*" can fill in many of the details of opening and closing the freezer and the can in the "measure coffee" routine in (10). This simplifies the stored schema, but of course then all the extra pieces have to be added online. If the coffee is not found in the expected place, the backup location is the store, triggering the construction of the whole trip-to-the-store routine.

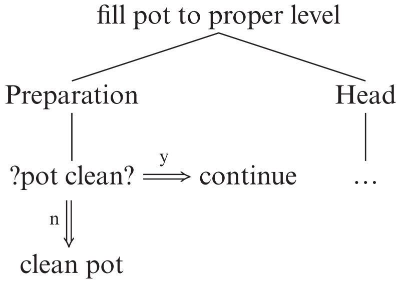
Next consider a problem posed earlier: suppose when you go to fill the coffeepot, you discover it needs to be cleaned. The cleaning routine itself oughtn't be a part of coffee-making: one cleans all kinds of things (and the motor control involved is phenomenally complicated—try to pay attention to your hands when you're washing dishes!). But when it is necessary to clean the pot, the cleaning routine has to be inserted as a Preparation for (9), perhaps as in (11).

(11)



One way to make this happen would be to add a checking step in the stored structure, perhaps like (12).

(12)



I'm not fond of this solution. If we put a checking step in the structure at every juncture where something might be needed or something might go wrong, we risk the computational explosion of the dreaded Frame Problem (McCarthy and Hayes 1969): for example, when walking from the coffeemaker to the sink, do you have to constantly check to make sure the sink is still there, the floor is still there, and your feet are still connected to your legs? There are too many ways things can go wrong and too many possible repairs to build them all into particular routines.

Another possibility for constructing (11) is that the coffee-making routine stipulates a clean pot, just as the shaking hands routine specifies an initial position for the hand. If the pot is already clean, no discrepancy is detected, so nothing needs to be added into the water-measuring routine. However, if the pot is not clean, "clean *x*" is called up as a Preparation to rectify the discrepancy.

But this solution still misses the point that *every* time we start to use a utensil, we want it to be clean. This suggests that there is a general constraint on the use of utensils that is unified with any particular food-preparation routine: it's not just the coffeepot, but *any* utensil that must be clean before use. Another such case might be the general routine that leads you to notice you're running low on coffee (or anything else) and thereby to write it on the shopping list—preparing a Preparation for an action that isn't even planned yet. Such general side-routines have something of the flavor of the "demons" of Lindsay and Norman (1977): they are, as it were, autonomously and unconsciously vigilant, ready to jump in whenever circumstances warrant. (Is obsessive-compulsive disorder a difficulty with such routines?)

Finally, let's think about Codas. I've kept talking about Codas as returning to the status quo ante. Here the discrepancy is between the state of affairs at the end of the Head of the action and the state obtaining before the action began. In the case of shaking hands, this Coda is motivated by necessity: you can't be attached to the other participant forever. In the case of the various steps of making coffee, the Coda is motivated more by attention and foresight, the desire to keep things in order (another demon?). So this sort of Coda is more dispensable, especially for children, whose attention and foresight are not up to their parents' standards. It seems to be a general problem in socializing one's children that they leave the Codas out of tasks. That's why we're always cleaning up after them and turning off the lights in rooms they left hours ago.

What I find interesting about this approach to action composition is the hypothesis that the main routine does not need to be responsible for calling all the subactions. "Making coffee" does not have to be full of checkpoints like "Is the pot clean?," "Is there coffee in the freezer?," "Is the top off the coffee can already?," "Am I at the sink yet?," and "Is the sink still there?" Rather, subactions are called by the integrative process that adapts the schema to a particular action in the present context. The integrative process is sensitive to discrepancies between the present situation and the requirements of the routine in progress. Moreover, having detected a discrepancy, the integrative process does not specify a particular subaction; it simply calls for any action that can rectify the discrepancy, and like words, suitable subactions "volunteer."

A more difficult case is when *no* suitable actions volunteer. This is perhaps the point where conscious planning has to step in, searching for a way to get to the desired point in a series of steps, each of which *is* already a stored schema. I'll not follow this line up, but here is where we make contact with other traditions in the planning literature, such as the old Newell and Simon (1972) General Problem Solver.

4.4.3 Choosing among Alternative Actions

There is a further parallel to lexical selection. Recall that when speech production calls the lexicon, all remotely appropriate words get activated, and then comes a process of selecting a single word as the one to be integrated into the present utterance. Parallel situations arise in constructing actions, often consciously (as word choice is, *very* occasionally). Consider a constituent of "making coffee" we haven't looked at in detail yet: "turn on machine."[7] Suppose I press the switch, but the machine doesn't go on. I may notice this because the little red light doesn't go on, or because 10 seconds later the water doesn't start hissing, or because two minutes later there still isn't coffee in the pot. For any of these cues to alert me, I need to know something about what to expect from the machine—particular knowledge about how the device works—and my attention needs to be drawn to the discrepancy.

Suppose I detect the discrepancy. What happens next? I have a huge variety of possible repair strategies in my action lexicon. Some might be very specific to this device: the coffeemaker might respond to a particular sort of banging on its top or jiggling its switch. Most are more general. For instance, among the strategies for dealing with an electrical appliance that doesn't turn on are checking to see if it is plugged in, checking whether a fuse has blown, and checking whether the power in the house is out. My wife says her most general strategy for attempting to repair *anything* is "call Ray."

[7] More particular knowledge of objects: you have to know where the switch is on the machine, and how to press it. Remember that these are not always obvious with a machine you've never encountered before—my favorites are the lamps in hotels.

One option for every repair routine is “abandon task.” If this option is chosen, abandonment percolates up the structure of the action to everything for which this subaction is necessary. So if you abandon trying to get the coffeemaker to turn on, you also abandon making coffee, and therefore (under the usual scenario) you also abandon drinking coffee. This is the counterpart in action logic of the propositional inference rule $[[P \rightarrow Q] \,\&\, {\sim}Q] \rightarrow {\sim}P$. On the other hand, alternative courses of action can be pursued further up the action tree. If the coffeemaker doesn’t work, one can still drink coffee by going out to a café.

How to choose which of the many possible repair routines to try? An old-fashioned programming approach to this plethora of strategies would be to have an ordered checklist. Sometimes we do use such a structured metastrategy. But I don’t think that’s a sufficiently general solution; there’s no reason to always try everything in the same order. Moreover, there are situations in which no mere list of possible repairs will do the trick. Suppose you and an acquaintance spot each other at a party, and it’s appropriate to shake hands. But she’s holding a plate and a glass, or you’ve been eating chicken wings so your hands are covered with grease, or she’s way across the room talking to someone else. On such an occasion you may improvise a symbolic handshake, say reaching in her direction and shaking your hand rigidly but without grasping hers. You may have never done this before; it’s a new action, created on the spot. So it can’t be on a list. (On the other hand, you may then store it away in a list for future such emergencies.)

In the spirit of the approach to the free composition of actions we’re trying out here, let’s consider another possibility. Suppose that, like multiple words “volunteering” to fit a desired meaning, multiple actions can “volunteer” to fill a gap in action—in fact, all reasonably appropriate actions do so (including modified versions of the normal routines such as the symbolic handshake). Returning to the coffee example, in response to detection of the discrepancy “switch pressed but coffeemaker hasn’t gone on,” a lot of different actions with different levels of generality are activated.

Which option is selected? Following the same intuitions that motivate the literature on rational choice, optimization, and heuristics, let’s guess that the primary criteria for selection are predicted benefit and predicted cost. If my coffeemaker isn’t working, checking the plug doesn’t cost very much, checking the fuse costs a good deal more, taking the coffeemaker apart a whole lot more. Similarly, in searching for one’s keys, one tries the most likely places first, in the hope of minimizing cost. Notice that relative cost has to be computed in a situation-specific fashion. For instance, the cost of abandoning attempts to get the coffeemaker to work depends on how much you want coffee (or think you'll want it soon). If making coffee is a Preparation for the "hospitality" routine, the cost is higher than if the coffee is just for yourself.

All of this presumes that you can estimate an action's cost in advance of performing it. My impression from a passing acquaintance with the literature is that the computations rapidly get fairly horrendous (though perhaps no more so than in lexical selection), and one frequently resorts to heuristics such as Gigerenzer et al.'s (2000) "fast and frugal strategies." Here is a place where even morality will impinge on the theory of action: moral judgment places biases on the value or cost of certain actions and therefore affects the choice among alternatives. We return to this issue in chapter 9.

4.5 Summing Up

What have we got out of this exercise?

- Even the simplest and most routine actions reveal a complex hierarchical structure, some of which is stored in memory as action schemas and some of which is the result of composing stored action schemas online.
- Some aspects of a stored action schema can be applied either to perceiving others performing the action or to executing one's own action.
- The structure of an action can involve both social and physical planes. The physical plane can involve both functional description (what is to be accomplished) and more strictly physical description (what motions accomplish it). The latter is linked to the actual motor script that realizes the action as muscle activations guided by proprioception.
- The structural relations underlying the composition of actions include
  a. the combination of actions as Head, Preparation, and Coda of a larger action;
  b. concurrent Modulation of a Head;
  c. Modulation of a Process by checking, which can terminate the Process;
  d. temporally unordered Heads (e.g. measuring water and measuring coffee);
  e. simultaneous Heads (e.g. shaking hands and making eye contact);

  and probably other possibilities. These structural options can be thought of as constituting parts of the "grammar" of actions.

- One stores vast amounts of information about how various devices work, potentially at every level of specificity from very particular (the faucet in my kitchen) to very general (electrical appliances). Perhaps "naive physics" is one of the most general schemas in this class: "how physical objects and substances work."
- One also stores vast amounts of information about the canonical location of particular objects.
- Much of the online composition of action is not driven by explicit choices in stored routines: making all the choice points explicit would lead to computational explosion (the Frame Problem). Rather, stored actions are to some degree skeletal and schematic. The full complexity of executed actions is a consequence of
  a. composing multiple stored actions;
  b. instantiating and binding variables in the schemas to suit the current context, including composition with other actions. The variables include characters in the action and locations to be moved to.
- Composition is often motivated by a discrepancy between the current situation and the situation required to initiate an intended action. The discrepancy can be a matter of physical necessity (getting your hand to the right place or handling an object), a matter requiring perceptual attention (the dirty coffeepot), or a matter of foresight (running low on coffee).
- A discrepancy triggers a call to the action lexicon, which is answered promiscuously by all reasonably appropriate stored actions. The action actually executed is selected according to minimum cost, where cost is at this point a highly context-dependent wild card.

Some of the discussion here is reminiscent of Schank and Abelson's (1975) "scripts" and Minsky's (1975) "frames." These approaches proposed that one stores structured knowledge of complex actions, so that one doesn't have to build them up from primitives every time (using a General Problem Solver or the like). They were mostly couched in terms of what you need to know about actions in order to understand stories—for instance, what you need to know about restaurants so you know there's an implicit waiter in the narrative "Bill ordered a hamburger and it was burned so he didn't leave a big tip." Here, of course, we're interested in what you need to know about actions in order to *do* them. Still, it's plausible that in the end these different tasks call on the same kind of knowledge. For example, our question of what you have to store in the coffee-making routine might translate into asking exactly what verbal and pictorial information you have to put on the instruction sheet that comes with the coffeemaker. (And I believe these earlier approaches recognized the potential convergence.)

My impression is that the script/frame idea foundered for a number of reasons: first, there were just too many contingencies that couldn't be mentioned explicitly in the script; second, it proved difficult to characterize situations that involved mixed scripts (e.g. a birthday party at a restaurant); third, there was no notion of inheritance hierarchy that permitted a smooth transition from very explicit to very general scripts. The approach described here attempts to avoid these problems by (a) invoking compositionality, in particular compositionality that is not always driven by explicit instructions in the schemas themselves, and (b) organizing action schemas in terms of inheritance hierarchies of generality. Here I have drawn on analogies to lexical access and composition in speech production, an area of cognitive science that has blossomed in the past 25 years. (The closest analogy in the AI research of the 1970s was perhaps Lindsay and Norman's (1977) notion of "demons"; Minsky's (1986) "Society of Mind" is along similar lines.)

As in language, there is the question of learning: how do you get all these schematic actions in your head, so you have this repertoire? Some of them might be explicitly taught, but others aren't. Where do such new items come from? They have to be constructed from pieces assembled online. Where do *those* pieces come from? As in the case of language, it might make sense to look for a primitive basis of features in terms of which actions can be assembled. This has to include not just basic actions but also the structural principles for connecting them, such as the notions of decomposition into Head, Preparation, and Coda. So we are looking at the prospect of a sort of "Universal Grammar" that is the initial state for acquiring an action lexicon.

I haven't dealt at all with the process of *consciously* constructing complex plans, where you're starting more or less from scratch, and many, many steps have to be put together for an action to work—say *inventing* a coffeemaker. Here I've been concerned with the more or less automatic construction of actions that we perform all the time in our daily lives. The point is to see how rich even these dumb unnoticed actions are. Of course, the more of these schemas you have in memory, the smoother your action becomes: a call to the action lexicon produces lots of useful results among which you can select, rather than *no* useful results, in which case you must rely on conscious ingenuity.

Let me conclude by summing up the parallels to the theory of language. Action structures, like linguistic structures, can be full of embedded constituent structure. Like linguistic structures, they seem to be determined by some sort of a grammar that specifies the structural options. Thus the grammar of action is a counterexample to Hauser, Chomsky, and Fitch's (2002) hypothesis to the effect that the presence of recursion is what makes human language unique among cognitive capacities. Furthermore, the action lexicon, like the linguistic lexicon, appears to be enormous, and some action schemas are partly indexed by the objects involved in the actions. Like the linguistic lexicon, the action lexicon is structured in terms of inheritance hierarchies, which relate very general to very specific schemas with all degrees of generality in between. Like the composition of linguistic structures, the composition of action structures involves instantiation and binding of variables, and possibly even quantification. We have also been able to make some plausible claims about the online construction of actions by taking seriously the analogy with language production. Moreover, like linguistic structures, action structures can be used not only to produce actions of one's own but also to understand the actions of others.

Those who wish to deny language much in the way of special character might therefore be tempted to say that we have shown language to be simply an outgrowth of action in general. I think such a conclusion would be misguided. First of all, at this point the theory of action is little more than armchair speculation, backed up by some research in planning and robotics but little psychological experimentation and no depth of coverage. So it hardly can serve as a serious justification for repudiating decades of linguistic research.

In addition, even at the primitive stage achieved here, it is possible to recognize important differences. To be sure, language partakes of many general principles of structural organization, memory, and processing that are shared by other systems of mind. That should be no surprise. However, one thing that makes language different is its role in the architecture of the mind, as a bidirectional conduit between the structure of thought and overt communicative expression. Action has quite a different role. Another thing that makes language special is the particular *forms of structure* that the general principles apply to. The tree structures for actions have different categories and arrangements of constituents from the tree structures for language, especially those for phonology and syntax (see chapters 1 and 2, also Pinker and Jackendoff 2005, Jackendoff and Pinker 2005). If there is any structural common ground between action and language, it is in conceptual structure, which presumably *can* encode aspects of complex actions. But conceptual structure is the organization of thought in general, and is not particular to language.

So the conclusion is mixed: The wondrous recursive creativity of language is not as special as it is often claimed to be. Nevertheless, language *is* a special system because of what it does and the particular structural materials it uses to do it.

Chapter 5

Cognition of Society and Culture

5.1 Social Cognition as a Cognitive Capacity

An important domain of human nature is *social cognition*, our ability to understand and engage in social interactions in the context of culture and social institutions.[1] This domain can be investigated in the larger biological context of how all sorts of organisms deal with conspecifics and how they understand the interactions of other conspecifics. Framed this way, social cognition can be investigated not only in terms of humans but in terms of animal societies as well, all the way from chimps to ants. This chapter lays out an overview of the inquiry. The first half of the chapter sets the stage, situating the enterprise in cognitive science and social science. The second half discusses a range of social and cultural phenomena that play a central role in the approach, some of which are taken up in more detail in part II of the book.

[1] In the middle 1980s, when I began thinking about the issues in this chapter, most work in this area had been done by primatologists and a few renegade anthropologists and biologists. Thus my earlier essays on social cognition (Jackendoff 1992a, chaps. 3 and 4; 1994, chap. 15) presented it mainly as a prospect for the future. The future has now arrived, and there is a flood of literature from psychology, neuroscience, child development, evolutionary theory, and cognitive anthropology, among other areas, to the extent that one can hardly aspire to do it all justice.

The notion of social cognition I am addressing here is a good deal broader than one that has achieved some currency in practice. For example, Decety and Sommerville (2003) take the subject matter of social cognitive neuroscience to be the way one represents the self and others, and how this leads one to identify with others. Gallese, Keysers, and Rizzolatti (2004) adopt a related stance, taking social cognition to be theory of mind and attributing it largely to mirror neurons. As will be seen, identifying other persons and understanding their beliefs, intentions, and emotions is only one aspect of the story.

Some readers may wonder: Why should we have to worry at all about a cognitive capacity for social interaction? Aren't all our interactions with others just determined by (or constructed by) culture? The reply is that in order for an organism to interact with others, it has to have a mind/brain. Rocks and trees don't have social interactions. And fish and cats and even chimps don't have the same kinds of social interactions we do. So, even if you insist that our social interactions are determined by culture, it's still important to ask two questions.

- What is the character of human social knowledge, such that it can be stored and processed in a human brain?
- What is it about human minds/brains that makes them susceptible to being influenced and shaped by human culture—and what is it about cat and chimp brains that makes them *not* susceptible, even when they are extensively exposed to human culture?

To simply attribute this difference to human plasticity is not enough. Over the past few decades, it has become clear that human brains are not equipotential blank slates, ready to take up whatever the environment happens to present them with (Pinker 2002). Moreover, it has also become clear that a capacity to learn *is* a cognitive capacity, not just the absence of a rigid instinct. The evolutionary transition from ape to human cognition is to be characterized not as a loss of instincts, but as a gain in ways to learn.

Just to provide a point of comparison: many people still hold the view that *language* is entirely a cultural artifact. However, the last 40 years have shown tremendous dividends from studying what it is about the human mind/brain that makes it possible to master language (see chapter 2). The study of social cognition explores a similar approach to society and culture. Just as linguistics has come to focus on the knowledge and ability of the individual language user more than on "the English/Spanish/Japanese language," so a theory of social cognition focuses on the individual's knowledge, understanding, and ability in social/cultural contexts more than on the structure of the culture as a whole.

Still, why should we want to separate social cognition from other, more general cognitive processes? A first answer is that other sorts of knowledge such as language, number, "naive physics," and "naive biology" have been profitably studied as specialized systems of mind. On the face of it, social cognition presents itself as another such content domain, a potential "core domain of knowledge" in the sense of Spelke 2003. A second answer might point out that many subareas of social cognition have already been examined in terms of specialized capacities: sexual selection, face recognition, altruism, cooperation, morality, theory of mind, cheater detection, and so on. So considering social cognition as a whole is actually a move toward unification rather than one of separation.

A third answer is that treating a mental domain as separate does not require isolating it from the rest of the mind. For instance, language would certainly be useless without its connections to thought, perception, and action; and a sense of number would be useless without a conception of things to count. Similarly, the brain and the heart are deeply interdependent, but that does not stop us from recognizing that they are separate organs. So if social cognition is to be described as a separate mental capacity, then the description naturally has to include its interactions with other capacities.

In order for an organism to make use of its capacity for social cognition, of course it has to be able to perceive the environment and act in it. I take it, though, that a theory of social cognition can abstract away from most problems of basic perception and motor control and can concentrate on phenomena that are more strictly of social significance. Social significance does penetrate into motor control in the production of social signals such as facial expressions and communicative calls, not to mention production of language. And perception has social significance in phenomena such as face perception, the tracking of eye gaze, and recognition of affect through facial expression, gesture, posture, and motion: all of these fall under "person perception." Such perception is surprisingly subtle: it includes phenomena studied in the often-cited work of Heider and Simmel (1944) and more recent elaborations such as the work of Bloom and Veres (1999), in which subjects cannot help attributing intentions to simple shapes such as triangles on the basis of the character of their motion.

These motor and perceptual phenomena are by all means fascinating, and their study goes back to Darwin. For my purposes here, though, the more important question for a theory of social cognition is what purposes these perceptual and motor phenomena serve. Why should the organism care so much about who it's seeing and what sorts of interactions to expect from that individual? To answer these questions, we need to think about what *cognitive* systems of mental organization make face and affect recognition and the production of signals crucial to the organism.

5.2 Parallels with Linguistics

This is where I come in as a linguist. The problem of how a child acquires social/cultural competence bears a deep analogy to the problem of language acquisition, which, as observed in chapter 2, has formed the foundation of contemporary linguistic theory for the last 40 years. It is worth reviewing this problem, recapitulating the discussion of section 2.2, before translating it into terms appropriate for social cognition.[2]

First, as just mentioned, contemporary linguistics is mentalistic, focusing on the character of the individual language user's cognitive capacity. Through this cognitive capacity, humans manage to create and understand an unlimited number of utterances of their language, most of which they have never heard before. The ability to use language must therefore involve a combinatorial system of principles (or a grammar) in the language user's mind/brain, which allow linguistic structures to be constructed from some finite stock of learned elements stored in memory. The grammar is not available to the consciousness of the language user; only its output is available.

The child must acquire this system in the course of learning to speak. The child has no direct evidence for the grammar: again, only its output is available, in the speech of those with whom the child interacts. Learning therefore must involve the active creation of organization in the mind/brain of the learner; it may or may not involve active teaching on the part of those with whom the learner interacts. In order to use speech

[2] It is not so new to draw a parallel between social interaction and language. Rawls (1971) pointed out that we can make intuitive judgments of the justice or fairness of a situation, often in instances where we cannot explicitly state the principles behind our judgment. He drew an analogy to our ability to make judgments of the grammaticality of sentences, citing Chomsky's then-new theory of grammar. This parallel has been picked up more recently by Mikhail (forthcoming) and Hauser (2006), who explore theories of morality modeled on linguistic theory.

I think Rawls was on the right track in seeking a mentalistic theory of justice. However, like many people in many different fields at the time, he drew too shallow a parallel between linguistics and his own concerns. On one hand, he tended to overlook cultural differences in moral judgments (see section 5.10), which would have strengthened the parallel to linguistics. But on the other hand, the parallel as he drew it is not entirely apt. The basic point of linguistic competence is not the ability to make grammaticality judgments; this is merely a side effect of being able to use language for communication. In contrast, one's ability to make judgments of justice or morality *is* the point of moral competence: it helps determine how one behaves toward the person being judged, including oneself (see chapter 9). Here I propose a deeper starting point for the analogy, going back to the first principles that motivate the theory of generative grammar. A closer antecedent for my approach is Macnamara (1991), who draws a parallel between moral reasoning and intuitive geometry.

in the environment as evidence for the grammar, the child must bring to bear inner resources of the mind/brain. Since these inner resources are by definition not learned, they must be a consequence of the inherent structure of human brains, determined by the interaction of the genome with the processes of biological development. Some of these inner resources may be cognitive specializations for language in particular (the so-called narrow faculty of language); others may be applicable for purposes more general than just learning language. In principle, it should be possible to sort these out. So goes the argument that grounds modern research into language.

We can state an almost parallel suite of issues for social cognition. The *answers* may or may not turn out to be parallel, but the questions are surely legitimate. As suggested in the previous section, the focus of the inquiry is mentalistic. The basic observation is that humans manage to participate in and understand an unlimited number of social interactions, most of which they have never encountered before in exactly the same form. The ability to interact socially must therefore involve a combinatorial system of principles in each individual's mind/brain, which make it possible to build up understanding of particular situations from some finite stock of stored elements. Social cognition differs from language in that there are certainly some principles that individuals can state explicitly. But, as we will see, there are also principles of interaction and social understanding that, like the f-rules of language (to use the term of chapter 2), are quite inaccessible to consciousness (i.e. "intuitive"). If so, it's of interest to ask how one's understanding of social interaction is parceled out between explicit and unconscious aspects.

Whether the principles are conscious, unconscious, or some mixture, the child must acquire them in the course of being socialized. For the conscious parts, the child often gets a lot of explicit teaching from caretakers and even older children. For the unconscious parts, though, the only thing the child has to go on is examples of actual social behavior, without explicit interpretation. This means that the child must be actively creating interpretations that lead to acquiring principles of social interaction. In order to accomplish this, the child cannot be flying blind: there must be inner resources in the child's brain that make this learning possible. Since these inner resources are by definition not learned, they must be a consequence of the inherent structure of human brains, determined by the interaction of the genome with the processes of biological development. Some of these inner resources may be a cognitive specialization for social interaction; some may be applicable to purposes other than learning a

Table 5.1
Parallels between language and social cognition

• Unlimited number of understandable sentences	• Unlimited number of understandable social situations
• Requires combinatorial rule system in mind of language user	• Requires combinatorial rule system in mind of social participant
• Rule system not available to consciousness	• Rule system only partly available to consciousness
• Rule system must be acquired by child with only imperfect evidence in environment, virtually no teaching	• Rule system must be acquired by child with only imperfect evidence, only partially taught
• Learning thus requires inner unlearned resources, perhaps partly specific to language	• Learning thus requires inner unlearned resources, perhaps partly specific to social cognition
• Inner resources must be determined by genome interacting with processes of biological development	• Inner resources must be determined by genome interacting with processes of biological development

social system. In principle, it should be possible to sort these out. Table 5.1 summarizes the parallels.

A further parallel, of course, is that both language and culture depend on the existence of a community for their functioning and transmission. In order for both the language and the society as a whole to work properly and remain stable, individuals making use of the language or participating in the culture need to have essentially the same cognitive organization, with some tolerance for individual variation. It is the stability and relative uniformity of the system that gives rise to the impression that the language and the culture are independent objective entities that transcend their individual participants. Moreover, the transmission of language and that of culture follow similar lines at the level of entire societies over historical time; the study of this parallel is now a flourishing enterprise (Cavalli-Sforza 2001). (We return to this impression of the objectivity of culture in chapters 7 and 9.)

If we follow this line of inquiry in parallel with linguistics, all aspects of cognitive neuroscience come to bear on the problem of social cognition. Paralleling linguistic theory, we can study the structure of social understanding and look for universals and statistical tendencies of human culture, using the tools of anthropology. Paralleling neurolinguistics and psycholinguistics, we can ask about the neurological and genetic bases of social cognition, and about the cognitive processes of social cognition, that is, how the brain processes, accesses, and stores social information in real time. Paralleling developmental psycholinguistics, we can ask about the course of the child's social development. In short, all the angles available for studying the language capacity have analogues in the capacity for social interaction. And in fact all of these approaches are by now amply attested in the literature.

The value of the parallel between innatist approaches to language and to social cognition depends, of course, on what the social behavior of humans proves to be like. Is it comparable in complexity to language, or does it just consist of relatively simple patterns, "habits," "memories," or "templates"? If the latter, then the opening premise of my argument, namely that humans participate in and understand an unlimited number of different social interactions, is false, and the rest of the argument reduces to relative triviality.[3] However, to the degree that social behavior is complex and subtle, the parallel goes through and the rest of the issues are indeed of interest.

In fact, to assert that social behavior is just a set of "habits," "memories," or "templates" is simply shrugging away the problem. When one attempts to formulate the content of a "habit" in detail and show how it is applied in a wide variety of circumstances, one is led ineluctably to the notion of a piece of cognitive structure or action structure containing one or more variables (or "slots") that can be satisfied by details of the current situation. In the relevant respects, this is absolutely parallel to the formulation of principles of language advocated in chapter 2: flexible behavior is the result of combining structures stored in memory, where each structure contains factors that can be adjusted to context. What the theorist *calls* these structures—"rules" or "habits"—is a matter of personal choice, but their formal organization and interaction is the same in any case.

Of course, many people still think of language as a collection of simple "habits" too, in ignorance of the sorts of complexity represented in figure 1.1 for a trivially simple sentence (see sections 1.2 and 1.3). I suspect that those who are skeptical about the indefinite variability of social interactions are not looking at a fine enough grain of interaction. For instance,

[3] This argument against my approach is not just a straw man. It was offered to me in all seriousness by a prominent cognitive anthropologist who nevertheless is a devout Chomskyan when it comes to language.

the opening of Searle 1995 describes the richness of social understanding that lies behind the simple act of ordering a beer in a café. The even simpler case of shaking hands was discussed in section 4.2. This social interaction seems on the face of it totally stereotyped; yet we have seen that even here there is considerable subtlety—all of it normally unconscious. Traditions of "thick" description in anthropology and sociology (Berger and Luckmann 1966; Geertz 1973; Goffman 1974; Eibl-Eibesfeldt 1989) amply demonstrate how this complexity scales up phenomenally when dealing with real social life.

An additional line of evidence is available for human social cognition that is not available for language: comparative ethology. Social organization shows considerable variation across species. In particular, primate societies are highly structured and vary from species to species along dimensions such as characteristic size of social group, monogamous versus harem-based versus polygynous relations between the sexes, characteristic modes of aggression and reconciliation, and the character of dominance hierarchies (Smuts et al. 1987). This hints at a strong innate genetic basis to their social organization. Moreover, much of primate social behavior looks quite familiar to us, involving issues of kinship, dominance, alliance, group membership, and reciprocity (among many others, Goodall 1971; Smuts et al. 1987; Byrne and Whiten 1988; Cheney and Seyfarth 1990; de Waal 1996). This suggests (following Darwin) that behind human culture lies a firm foundation of primate evolutionary ancestry.

Within each primate species, especially apes, there is a certain amount of variation from one population to the next (Whiten et al. 2001; van Schaik et al. 2003). I'm not too interested in whether the variation among chimpanzee and orangutan populations should be called culture or not; everybody acknowledges that it's vastly more constrained than that among human cultures. Given the vast range of variation in human social behavior and social organization, learning plays a much more significant role in the socialization of humans than it does with other species.

The issue for the evolution of social cognition, then, is not just what problems early hominids had to face (as stressed by Tooby and Cosmides (1992), for instance), but also what problems earlier ancestral primates (and mammals before them) faced as well, and therefore on what prior solutions hominids were able to build. Using standard comparative methods, we can form hypotheses about the ancestral great ape social repertoire, and we can ask what tricks evolution had to add to the ancestral repertoire to get modern chimps, bonobos, gorillas—and us.[4] In addition to language, the literature has considered relevant innovations in the human lineage such as facility at imitation, ability to understand pointing, highly developed theory of mind, ability to engage in large-scale cooperation, and possibly even deep understanding of physical causality (Povinelli 2000; Tomasello 2000; Boyd and Richerson 2005).[5]

A further dimension has not played much of a role in the theory of language, but it could—again with a parallel in the theory of social cognition. It may be that some linguistic and social phenomena are not strictly speaking part of the abilities of the individuals taking part in them, but are rather an emergent property of individual behaviors in concert, plus environmental contingencies. For instance, cooperative behavior among ants is probably not the product of overt agreement among them or some fancy theory of mind. Rather, it's just the consequence of a number of "cheap tricks"—automatic responses to particular actions by conspecifics such as emitting smells. I don't care whether we want to call this collective behavior social cognition or not, but it's not without interest to explain how the social dynamics of ant colonies arise through the interaction of relatively stupid individual agents who are not gauging the overall consequences of their actions.

[4] A caveat: evolutionary psychology is a valuable tool, but only one of many in this enterprise. Although evolutionary considerations are important, I don't think every aspect of human social organization can be explained, much less discovered, by appeal to first principles of evolutionary psychology. In the case of language, little of its structure can be predicted solely on evolutionary first principles (Jackendoff 2002a, chap. 8).

[5] Boyd and Richerson (2005) make important connections among some of these capacities. They argue that convergence among the social cognitive structures in a human community, necessary for both culture and language, takes place only because humans have the ability—and compulsion—to imitate those with whom they interact. The result is a tendency toward mutual "tuning" of behavior within the community. Moreover, they argue, using computer modeling, that the compulsion to imitate is a necessary component in order for large-scale cooperation to develop. Thus they see a major divide between humans and ape societies, the extent of large-scale cooperation, as having arisen from a more basic major difference in the ability to imitate.

Alexander (1987) takes a complementary tack. The motivating fact for him is that humans' main predator is other humans. Projecting this back into hominid societies, he argues that the advantage of large-scale cooperation was better protection for the group against other groups.

It might be that some aspects of human cultures have this character too. For example, the behavior of the stock market may be like this: usually its gyrations are not the direct effect of planning, but instead emerge from the behavior of many independent agents. Many aspects of language change (such as the change from Shakespearean to modern English) are also attributable to such effects. Donald (1998) claims that a great deal of human social behavior, including language, is of this character; I'll grant him that some of it is. The point is that we should be careful not to overinterpret human social behavior in terms of highly structured cognition.

Still, we shouldn't underinterpret it either. The example of linguistics again comes to mind. Linguistic behavior, at all levels from phonetics to pragmatics, has proven far more highly structured than anyone would have expected 50 years ago. It's just that most of the complexity of language falls below the radar—it is transparent to us. It is only with training in linguistics that we begin to notice it. I suspect the same is true of social behavior. The trick is to figure out *exactly* what capability should be ascribed to the individual, what part is due to group dynamics, and how the two together result in the observed complexity of structure. The part due to group dynamics corresponds to the traditional sense of "social construction" of culture: it emerges out of the collective practices of the group, without anyone necessarily intending it to do so.

Eventually, the individual's capacity for social cognition has to be explained in terms of the brain and the genome. However, perhaps betraying my training as a linguist, I'm most interested in social cognition for the moment from a formal or functional point of view. At our present stage of understanding the brain, we may be able to localize some social function, say face recognition in the right parietal lobe or moral judgment in prefrontal areas. Or we may be able to find a neurotransmitter that enhances aggression or affiliative behavior and pinpoint its locus of action in the brain. But figuring out how the whole system works is probably better pursued at a level of abstraction somewhat distant from the neurons. Again, I see this as analogous to the study of language. As emphasized several times in previous chapters, we know a fair amount about localization of different aspects of language function in the brain, but it's likely to be a long time before we understand how the details of Finnish case marking and Hausa tone in language—or the notion of ownership in social cognition—are instantiated in the brain. At present, it seems more fruitful to look at these problems from the perspective of an abstract structural grammar or internal logic and to put off issues of neural instantiation for a while. This doesn't preclude applying the tools of neuroscience by any means—we have already learned a lot there as well. But the two efforts ought to run in parallel.

5.3 Objections from Social Science

This cognitive, biological, and evolutionary approach is not the predominant way that culture is studied. At least in America, anthropology and sociology are dominated by the view that humans are totally a product of their culture and that it is meaningless to claim that culture is the way it is in part because of human cognitive abilities. Tooby and Cosmides (1992), Ehrenreich and McIntosh (1995), and Pinker (2002) document how widespread, influential, and deeply entrenched such attitudes are, not only among scientists but also in politics and ordinary life.

There were originally good reasons for such attitudes. Degler (1991) shows that in the late nineteenth and early twentieth centuries, it was widely assumed (even by Darwin, Humboldt, and the distinguished early twentieth-century linguist Otto Jespersen) that races and ethnicities are sharply distinguished in intelligence and moral capacities—with northern Europeans naturally at the peak of both.[6] Such views comported with and reinforced the rampant colonialism of the period, and in the United States they were invoked to justify ethnically specific anti-immigration legislation. Degler describes how the anthropologist and linguist Franz Boas fought fiercely for cultural and linguistic relativism in the interests of resisting such discrimination. The well-known work of his students Margaret Mead and Ruth Benedict was intended to demonstrate the unlimited variability and equal value of cultural institutions and moral systems across the world. And Boas's views won the day in anthropology, sociology, and linguistics.

The outcome has been a downplaying of the cognitive aspects of culture in favor of the environmental: "While it is possible to say that man

[6] A recurrent mistake, so common that it's tempting to attribute it to human nature, is to draw a parallel between human races and animal species in their degree of "advancement" (Hirschfeld 1996). Even E. O. Wilson, a pioneer of the cognitive and evolutionary approach to social organization, succumbs to this in his *Sociobiology* (1975). The whole book is about innate species-specific differences that govern social behavior. Then suddenly, in the final chapter on humans, there are passages about heritability of racial and cultural differences. Here I am interested in the characteristics of the species as a whole (as is Wilson, most of the time).

has a nature, it is more significant to say that man constructs his own nature, or more simply, that man produces himself" (Berger and Luckmann 1966, 49). However, the conclusion that culture is the wellspring of human nature is not inevitable. Even if one argues against racist interpretations of cultural differences in economic, technological, and military development,[7] one need not conclude that *no* aspects of human cultural capacity have an inherited basis and that there is *no* inherent human nature of relevance to the social sciences, invariant across the species.

It is still worse to conclude, as I gather many in anthropology have done, that it is impossible—and wrong—to attempt to go beyond a radically local, relativistic, and contextualized perspective on the culture one is studying (these attitudes are documented by Brown (1991), Ehrenreich and McIntosh (1997), and Zuriff (1998)). Such a stance renders impossible any sort of scientific crosscultural comparison. Some rejoice at this. I don't. Scientific paternalism and condescension toward other cultures are indeed objectionable—"We enlightened scientists know better, and aren't you exotic." This only mirrors the colonialist and imperialist attitudes of a century ago. But as with the description of other people's languages, I see no problem with doing the best, most honest job of description we can, respecting native intuitions and recognizing our fallibility. That's just good science. Moreover, efforts to carry out such crosscultural comparison, such as Brown 1991 and the monumental Eibl-Eibesfeldt 1989, reveal a rich tapestry of universal or at least widespread patterns of social behavior and organization.

A different critique of the biological/psychological/evolutionary approach (e.g. Sahlins 1976) takes the position that human cultures cannot be deterministically grounded in biology, because their coherence depends on systems of layered symbolic meaning. I cannot do full justice to this objection here. However, the following points can be made. First, addressing the "deterministic" aspect of the objection: the existence of an innate cognitive capacity for social and cultural interaction does not thereby imply that cultures should be uniformly structured. Rather, on

[7] This argument has been taken up again in our time by Jared Diamond (1997) against recent resurgences such as Herrnstein and Murray 1994. On the other hand, Diamond's argument that such differences are largely the result of ecological opportunities does not fully predict cultural outcome either. For instance, case studies documented by Boyd and Richerson (2005) and Atran, Medin, and Ross (2005) show that different cultures can exist side by side for a long time in the same physical environment, with different economic consequences. And of course language (apart from environment-specific vocabulary) is completely independent of physical environment.

the view advocated here, what is innate is a capacity to *learn* varied cultures from one's environment. Like the faculty of language, it allows vast variation; part of the empirical problem is to determine what the range of variation is. In other words, a faculty for social cognition can be thought of, not as a prescription of universals, but as a "toolkit" of issues that societies must address in one way or another. Through this toolkit, children learning a culture are alert to detect environmental cues for how the culture realizes these issues, and they are innately provided with some building blocks for constructing the relevant concepts.

Another answer to the charge of determinism is that cognitive processes characteristically involve a tension among conflicting principles. A clear instance in the social domain is sexual behavior. Primitive sexual drives ("Copulate with any individual of the appropriate sex") are overlaid with Darwinian principles of sexual selection ("Strive for partners that give your genes the best chances of survival"), which have different realizations from species to species. In humans, these are in turn overlaid with variable cultural principles of courtship, contractual marriage, and so on, which arise from the social/cultural capacity. For example, the incest taboo has a grounding in Darwinian principles of sexual selection, but it is overlaid with cultural elaboration and can itself be undermined on occasion by primitive sexual drives. Thus there is no reason to expect a biologically grounded cultural capacity to produce uniform and deterministic behavior.

Let me finally address the assertion that culture involves symbolic meaning and therefore cannot be biologically grounded. Again consider language: language is nothing if not the symbolic expression of thought. Although each particular language is a product of cultural transmission, each speaker learns it by virtue of an innate disposition to interpret the noises made by people in the environment as symbols built out of phonemes and structured as nouns and verbs. Similarly, the symbolic interpretation that one puts on the world by virtue of being a member of a culture is the product of a cognitive capacity that inclines a culture learner to seek such symbolic interpretations. Moreover, similar issues occur over and over again in the content of these interpretations, combined and recombined in ever-varying fashion. Among these issues are the intentions and goals of others, kinship, group membership, dominance, alliance, friendship, enmity, rights and obligations, and the relation of humans to the natural and supernatural world. Such concepts do not come for free in the cognitive repertoire. Cows don't have them, and chimps have only some of them, and these only in a limited way. So ultimately we come around yet again to the question of what it is about humans that permits such concepts to structure our perception of the world and our action in it. This is the inquiry being undertaken here.

There is no question that anthropological description and interpretation have much to offer in terms of data about social behavior and social organization, particularly in documenting details of crosscultural similarity and variation. But a theory of social cognition becomes quite different when we take seriously the cognitive capacity behind the overt phenomena that anthropologists study.

5.4 A Role for Linguistics

Beyond the parallelisms between language and social cognition laid out in section 5.2, there is another reason for linguists to be interested in social cognition: some of the most interesting and problematic issues in syntax and semantics involve predicates in the social domain.

For instance, consider the concepts expressed by *X requested Y to do such-and-such* and *X ordered Y to do such-and-such*. Both involve X making some utterance, intended to be heard by Y, that concerns X's wish for Y to perform some action. What is the difference between them? The main difference appears to be that, with *order*, X is in a position of social dominance over Y and therefore can invoke the authority to impose sanctions on Y if Y does not perform as X desires. In addition, both X and Y have to be aware that this relative position obtains. Something misfires if X issues an order that Y recognizes only as a request, or if X issues a request that Y interprets as an order. In short, an order is something like a request backed up by the conventions of social dominance. Note that the social dominance must be *invoked*: dominant individuals can make requests as well as issue orders. (A similar analysis, with much more comprehensive detail, appears in Bach and Harnish 1979.)

Crucial to this analysis are not only the social dominance hierarchy and awareness of it, but also the consequences: X's authority to issue orders, to expect compliance, and to impose sanctions if not obeyed (see chapter 11). As will be seen, this conjunction of factors recurs in a variety of situations, suggesting that it is a basic part of the logic of social cognition.

Next consider verbs that express transactions, for instance *buy*, *sell*, *rent*, and *trade*, staples of the syntactic literature for decades (Fillmore 1965; Miller and Johnson-Laird 1976, sec. 9.2 and references therein; Pinker 1989; Jackendoff 1990; Goldberg 1995). They all share the same basic semantic structure. Take a sentence like *Bill traded his bike to Harry for a horse*. Two actions are taking place: the bike is changing possession from Bill to Harry, and the horse is changing possession in the opposite direction, from Harry to Bill. However, there is a relation between these two actions. What makes them constitute a *trade*, rather than two unrelated changes of possession, is (for a first approximation) that Bill and Harry agree that the bike and the horse are of equivalent value, and that Bill and Harry acknowledge the paired changes of possession as linked by this equivalence. (More details in chapter 10.)

Apart from the actual physical transfer of the goods, every part of this involves a social concept. The notion of agreement—mutual acknowledgment and validation of the other's point of view—is a social transaction (see sections 5.8 and 8.8). The notion of the value of an object is not a physical attribute but depends heavily on cultural conventions (chapter 9). Moreover, behind the cultural conventions lies a more basic and very abstract conceptualization: the idea that incommensurate objects and actions can be reduced to a linear scale of equivalence along a dimension of value (chapter 9).

Nor is the pairing of changes of possession physically necessary for a trade to be concluded. It may still count as a trade if Bill gives his bike to Harry in exchange for a *promise* (or *obligation*) for Harry to turn the horse over to Bill at some later date. What is Harry's promise or obligation in this case? Roughly, it is a granting of authority to Bill to impose sanctions on Harry if Harry does not perform as promised. It is this notion of authority to impose sanctions (again), and Harry's willingness to accept these sanctions, that distinguishes a promise from a mere *prediction*. (See chapter 11 for obligations and authority. The logic of exchange is in fact still more complicated, as shown in Tooby and Cosmides 1989 and chapter 10; see also Bach and Harnish 1979 on promises versus predictions.)

In turn, for there to be a change in ownership in a trade, there has to be such a thing as ownership. What does *X owns Z* mean? Very roughly, following Miller and Johnson-Laird (1976), who in turn quote Snare (1972), the concept seems to have three parts:

A. X has the right (or authority) to use Z as he or she wishes.
B. X has the right (or authority) to control anyone else's use of Z and to impose sanctions for uses other than those he or she permits.
C. X has the right to give away rights A and B.

The primate cognitive heritage seems to offer two independent sources for ownership: territoriality and possession of food, each with its own special characteristics (see Eibl-Eibesfeldt 1989 for exploration of these antecedents).

For a slightly more complex case, *lending* something is allowing someone to use it, but without relinquishing ownership: the borrower is under obligation to give it back. In turn, *allowing* is granting a right, which presupposes the authority to grant it. That is, owning and lending make use of more of the same machinery.

Cultures differ in what one can own (artifacts, land, one's children, one's wife, slaves, or in our culture, rights to one's creative work), in what sanctions are imposed for misuse of someone else's property, and in how society imposes those sanctions. In addition, some cultures have notions of property rights for certain objects, particularly land, that involve only clauses A and B above, or only clause A (Eibl-Eibesfeldt 1989; Alan Fiske, pers. comm.): one can use land and perhaps prevent others from using it, but one does not have the right to give it away or sell it. But the basic framework is there universally—the child only has to learn what parameters govern ownership or property rights in the local culture.

Similarly, the fact that the social dominance hierarchy is invoked as part of the meaning of many predicates does not establish who is socially dominant in a given culture or how dominance is established. But without the basic notion of social dominance, all the actions that serve to instantiate a given culture's realization of it would be incomprehensible.

The story that emerges from these little examples is that even rather simple words of English such as *own*, *trade*, *promise*, *request*, and *order* reveal an underpinning of basic social concepts. Notions like dominance, authority, privilege, right, obligation, value, the imposing of sanctions, agreement, and so forth keep recurring as components of concepts in the social domain, just as notions like physical object, motion, location, and force pervade the domain of spatial concepts. These notions are abstract components out of which different cultures build different realizations. This gives a new kind of evidence for the existence of a specialized domain of social cognition; such evidence can be triangulated with evidence from anthropology, psychology, and primatology to give us some idea not just of the existence of the social capacity but of its content as well. Let us now turn to some of these content areas; chapters 6–11 explore several in more detail.

5.5 The Physical and the Social/Personal Domains

Where is social cognition localized in the functional ecology of the mind? From the discussion of the last section, a basic hypothesis emerges. Social cognition is one of the central systems of cognition. It is a major department of the level of conceptual structure (in the sense of chapter 2), standing alongside and interacting with another "core domain" (in the sense of Spelke 2003), the understanding of physical space.

The cognition of space involves concepts of physical objects that are located in three-dimensional isotropic space, that move in this space, and that exert forces on each other. Among the physical objects are natural objects like rocks and trees and rivers, artifacts with affordances for use like bicycles and tables, and animate objects like ants and worms and rats and tigers. The animates, unlike the rest, are conceptualized as capable of unpredictable self-initiated motion (i.e. volition)—and therefore, perhaps of desires, intentions, and even emotions as well. That is, animates are understood according to the *intentional stance* (in the sense of Dennett 1987).

The basic entities of the *social* domain are *persons*—individuals with whom we can have social relations. This domain encodes the relations and actions among them *as* persons: among other things, persons have social roles and responsibilities, and they are subject to moral judgment. It is our personhood that is taken to "raise us above the animals." That is, understanding persons goes beyond the intentional stance to what we might call the *personal* or *social stance*.

Like all concepts, the concept of person has a certain amount of leakage at the boundaries (for the ubiquity of "leakage" in categorization, see Jackendoff 1983, chaps. 5, 7, and 8; 2002a, sec. 11.6). Pets probably count as "honorary" persons, and so do personified animals in folk tales and cartoons. But the mosquito buzzing in your ear, though animate, certainly doesn't. As for leakage in the other direction (to raise the first of various depressing phenomena to be mentioned here), it is an all-too-common social tactic to characterize members of another social group as animals (say pigs, dogs, or monkeys) rather than as persons, and to use that as an excuse for condoning ruthless behavior toward them. They do not qualify for social relations, and so, as with mosquitoes, anything goes.

The social domain, in contrast with the physical domain, is not an isotropic space. In physical space, between any two objects that are not touching, there is an intermediate region of space that can be occupied by another object. Such a notion makes no sense in social space—there is no continuum of "intermediate spaces" between people. To the extent that the social domain has a notion of distance, it is "social distance," measured in terms of divisions of kinship, class, status, group membership, and degree of intimacy or alliance. Of course, there are notions of force, coercion, and constraint in the social domain, sometimes connected with physical force or the threat thereof, sometimes not.

Human beings are conceptualized as occupying both the physical and social domains. This duality is culturally widespread in folk conception as the division between body and soul or body and spirit (Jackendoff 1992a; Bloom 2004).

Personal identity invariably goes with the social domain, as several observations show. First, there is a culturally ubiquitous belief in supernatural entities such as spirits, ghosts, gods, and souls that survive death (Boyer 2001). All of these are beings who lack definite physical bodies, yet have social relations with people and with each other; hence they exist in the social domain, but not the physical. Second, we have no problem conceptualizing persons coming to inhabit different bodies through reincarnation, metamorphosis, or body-switching (consider how easy it is to understand movies like *Freaky Friday*). Third, in dreams we sometimes "know" a person is different from the one he or she looks like: "In my dream I was talking to Uncle Sol, but for some reason he looked like Milton Berle." Fourth, an individual suffering from Capgras syndrome (McKay, Langdon, and Coltheart 2005) will claim that his wife (or some other socially significant person) has been replaced by an imposter—a different person—who looks just like her. In each of these cases, whether through religious belief, fantasy, or delusion, an individual's personal identity—in the social domain—is cut loose from the identity of the physical body.

A different sort of observation that supports duality and separation of the physical and social domains is the discomfort that people feel with the idea of golems, humanoid computers, and the like—physical objects that suddenly sprout social identity or personhood. Such beings always play an unsettling role in folk culture, including our own ("Will computers get so smart that they'll take over the world?"). And there is a sort of converse of this, which I'll return to in section 5.11: people have extreme difficulty accepting and reasoning within a materialist philosophy of mind—thinking about persons as being defined only in physical terms ("How can we be just machines? That deprives us of our dignity and moral standing!"). The point is that conceptually there is a transcendental difference between the physical and the personal, one that is essential to our sense of ourselves as human beings, and one that is virtually impossible to erase.

The social plane does not contain only persons. It also contains the relations and actions among them, insofar as these are socially defined. Like souls, social actions are on their own unobservable. They become observable only through their linkages with the physical plane. (Searle (1995) describes this as the logical priority of "brute [i.e. physical] facts" over "institutional [i.e. social] facts.") Some physical actions such as eating and walking make sense on their own. But some, such as performing religious ceremonies and shaking hands (chapter 4), make sense only as instantiations of (or symbols of) actions on the social plane. Moreover, a purely physical action such as eating with other people or walking in the context of a ceremony can be correlated with a social significance and thereby become symbolic. Likewise, choices of costume or speech style can be used to signify social roles, but they don't constitute the roles they symbolize—even if people sometimes act as if they do (e.g. a military uniform or a flag standing for a country). Even social actions in relation to disembodied spirits involve physical actions; we call such actions magic or ritual. That is, physical actions become social because we construe them as such in terms of the social plane. This is another sense in which one can speak of the "social construction of reality" (Berger and Luckmann 1966) or the "construction of social reality" (Searle 1995).

It is not as though people consciously separate the physical and social planes. Indeed, they may act as though they're inextricable. The issue here, however, is how to "carve a natural joint" in our intuitive modes of thought.

This idea of entities and actions being simultaneously formulated and interpreted in two parallel linked planes, largely below the level of awareness, is actually not so unfamiliar in the theory of cognition. A commonplace example comes from Pustejovsky 1995. Compare a brick and a book. At the physical level, they have similar properties: you can lift them, move them around, stack them up, and so on. But because a book is conceptualized as containing *information*, it has a whole other set of properties as well: you can read it, copy it, analyze it, understand it, and so on. Crucially, in order to convey information, a physical instantiation is necessary. But not all instantiations of information are objects; for instance, a linguistic utterance is a physical sound, not an object, and it too can be perceived, copied, analyzed, and understood. Thus it makes sense analytically to think of a book as partaking of two different natures in parallel, each of which can exist without the other. But it is not just the book that subsists in the two planes. The action of *reading* a book also involves both planes at once: one is both moving one's eyes over the physical page and taking in information. Pustejovsky introduces the term *dot-object* for such dual objects and actions (for reasons having to do with his formal notation that need not concern us here).

A deeper (but not unrelated) example of a dot-object is language, which proceeds in two parallel planes, phonology and meaning. One can associate different phonology with the same meaning by switching languages; one can dissociate sound from meaning (nonlinguistic sounds and nonsense syllables); and one can dissociate meaning from sound (i.e. thought). Meaning (at least others' meaning) is unobservable without being linked to speech. Yet this parallel organization is largely transparent to experience: it is common to think of meaning as just an inherent part of the spoken word. So it is, I think, with persons and their bodies.

Let us look at shaking hands again, a little more closely. As discussed in chapter 4, it serves as a mark of social connection, and the physical realization of this mark can vary from culture to culture. More precisely, the social meaning is apparently a display of mutual respect, an expression of good intentions toward the other, which carries with it a presumed inclination toward cooperation and trust. Similar greeting rituals, with similar social meanings, are attested in other primates as well (Watanabe and Smuts 2004). Now recall from chapter 4 that all but one of the five situations in which it is appropriate to shake hands are symmetrical: it doesn't matter who initiates the handshake. The exception is congratulation, where the handshake must be initiated by the congratulator. This stands to reason, in that congratulation, unlike greeting and taking leave, is inherently a display by the congratulator of respect for the congratulatee.

The assignment of social meaning to an action may vary from culture to culture and even from situation to situation. For example, consider the action of presenting someone with a gift. In addition to its physical aspect, the action carries the social meaning of transferring possession (i.e. exercising right C of ownership in the analysis above). But there may be a further layer: Is this gift to be construed as tribute, that is, as a display of respect by a subordinate individual toward a dominant? Or is it to be construed as largesse, that is, as an expression by a dominant individual of benign dominance (or condescension)? Or is it to be construed as a display of affection and mutuality? Or is it part of an agreed-upon exchange? Much depends on the circumstances, and there is plenty of opportunity for misunderstanding. It is this sort of layering and ambiguity that creates "thick" social meaning in the sense of Geertz 1973.

It is important to keep the notion of the social domain distinct from the theory of mind, the ability of humans to attribute beliefs, desires, and intentions to others. In human social relations, we typically attribute a mental life to the persons with whom we interact. But theory of mind is broader: we do not hesitate to attribute desires and intentions to a tiger that is stalking an antelope. It is beside the point whether the tiger *really* has desires: *our* folk theory of mind attributes them anyway. That is, theory of mind extends beyond persons to other animate beings.

Conversely, not all aspects of social relations require a theory of mind: for a person to be a member of a certain clan and therefore to have certain obligations, it does not matter what we think that person believes or desires—it is just an objective social fact. Moreover, it makes sense to attribute some sort of social cognition to monkeys, who, according to much current thinking, lack theory of mind. So although theory of mind clearly plays an important role in human social/cultural cognition, these two aspects of cognition are not coextensive. I will continue to stress this distinction, for, as mentioned in note 1, much research in what is called social psychology and social cognition is concerned only with inferring others' attitudes, intentions, and goals. There's much more to social cognition than that.

5.6 Affiliations: Kinship, Alliances, Dominance

A very important part of social cognition is keeping track of your relationships to others. This section and the next discuss long-lasting relationships. Sections 5.8 and 5.9 then go on to relations among individuals that can shift from moment to moment as actors undertake different activities.

Perhaps the most obvious of the long-lasting relationships is kinship. In every culture, each individual is in a special relationship with his or her parents, children, spouse(s), and siblings. Many aspects of this relationship arise clearly from our mammalian heritage, in which the parents (or mother alone, depending on the species) must take care of the young for some period of time. Evolution has provided us, like other mammals, with patterns of perception and behavior that make this care possible and basically pleasurable.

Kin altruism extends beyond parent-child relationships to include siblings and potentially more distantly related kin. Theoretical models based on the "gene's-eye view" predict such relations: since kin share genetic material, acts done on behalf of kin can lead to proliferation of one's own genes to some degree. (For the mathematics of such relationships, see Hamilton 1964; see also discussion in Dawkins 1989.)

In human societies, kinship bonds are extended to more distant relatives as well as immediate family. Many cultures have elaborate customs, obligations, and rights associated with being in particular kin relations.

For example, every culture has an incest taboo, but its precise extent varies from culture to culture—with which extended kin sexual relations are forbidden and with which they are permitted or even encouraged.

Although we have perceptual cues for who is in our immediate family (the people we live with),[8] we don't have any such cue for more distant relatives. We rely on someone *telling* us we're related, and miraculously we come to feel the bonds of kinship. People easily can feel warmth toward distant cousins they have never heard of, upon meeting them for the first time at a family reunion. Likewise, an adopted child often feels affection for a newly discovered biological parent. Such examples show that the bond of kinship is not just *per*ceptually based, but also *con*ceptually. When we think of a person as kin, we feel and behave differently toward him or her.

A different sort of relation is that of allies or friends, unrelated individuals between whom there is a voluntary and lasting commitment to cooperative activity. The flip side is the relation of rivals or enemies, between whom there is a lasting commitment to competition. In both cases, participants know what they can count on from each other. My impression is that in many cultures such relationships can be formalized by oaths and the like—institutionalized agreements to establish the mutual relationship (see the discussion of agreement in section 5.8). Such relationships are documented in the primate literature as well (Goodall 1971; Smuts 1985).

More prominent in the ethological literature is discussion of dominance, a relation between two individuals whereby one (the subordinate one) regularly defers to the other (the dominant one) in matters of food choice, sexual selection, grooming partners, and so forth. Dominance is often based on size and aggressiveness, but it doesn't have to be. For instance, it can depend on kinship relations, as in vervet monkeys, where the children of highly ranked mothers inherit high rank (Cheney and Seyfarth 1990). This means that dominance cannot be a purely perceptually based relation: it too needs a conceptual basis.

In animal societies, dominance relations often fall into a linear order: if A is dominant to B, and B is dominant to C, then A is also dominant to C; and every individual in the group has a distinct place in the "pecking order." Dominance hierarchies characteristically remain stable over time, but subordinate individuals may mount challenges that, if successful,

[8] This seems to be the "cheap trick" that lies behind biologically based incest avoidance. See Eibl-Eibesfeldt 1989, sec. 4.6, for discussion of cases where "incest avoidance" has developed among unrelated individuals raised together.

rearrange the pattern. Cheney and Seyfarth (1990) present observational and experimental evidence that vervet monkeys know not just their own relations to all the other monkeys in the group, but also the relations of other monkeys to each other.

Dominance relations are pervasive in human societies too. But humans do not have a single pecking order; rather, dominance can be organized along many different dimensions, such as parent to child, teacher to student, boss to worker, ruler to subject, celebrity to fan, and of course in many cultures, husband to wife. It seems to me that when larger-scale human dominance hierarchies develop, they differ from animal hierarchies in tending to be pyramidal rather than linear: there is a top person dominant to a number of relatively equal subordinates, each of whom is dominant to further subordinates, and so on. This drastically expands the size of the group over which dominance can be extended. Still, the basic notion of a stable asymmetrical relationship based on deference of one individual toward another bears a strong resemblance to the animal model.

All of these relationships require that you keep track of who is who in your social milieu. Presumably this is the functional motivation behind the perceptual specializations for face and voice recognition.

5.7 Groups

5.7.1 The Axioms of Groups

Another kind of lasting affiliation, forming one of the most important elements of social structure, is group membership (stressed, for instance, by Alexander (1987), Eibl-Eibesfeldt (1989), Gilbert (1989), and Boyd and Richerson (2005)). The fundamental premise of the logic of groups is that some set of individuals constitutes a group, and everyone else is not a member. Typical examples are clubs, orchestras, and religious congregations. Families, extended families, and clans are particular sorts of groups that add kinship relations on top of the basic premise. Mere aggregations, such as the people who happen to be on the bus with me at the moment, don't constitute a group in this sense. Neither do "affinity groups" such as the baby boomer generation: although baby boomers might have similar goals, different from those of other generations, I don't think they feel any special "loyalty" to other baby boomers in the requisite sense.

The point of groups is that one's actions toward others can be conditioned not by who they are as individuals, but by whether or not they are members of one's group. The most basic principles of groups appear to be the following axioms. They pertain to primate groups, and to every kind of human group from teenage cliques to nations, with professions, religions, and social classes in between.

Axiom 1 Other things being equal, if you are a member of my group, I will behave favorably toward you. In particular, I will be willing to cooperate with you; and I expect the same from you.

Axiom 2 Other things being equal, if you are not a member of my group, I will behave unfavorably toward you. In particular, I will compete with you; and I expect the same from you.

Groups can differ in how great the disparity is between how one treats members of the group and how one treats others; this is part of the group's cultural mores. But the basic logic remains intact.

Given this logic, it's important to be able to determine who's in and who's out of one's group, especially when groups get so large that members are not necessarily acquainted with everyone else in the group. Members of human groups often make themselves more easily identified by adopting characteristic dress, customs, and manners of speaking.

In the human case (far more than with animals), one typically identifies with numerous overlapping and hierarchical groups. Should I act at any particular moment as an academic, a cognitive scientist, a linguist (one chain of embedded groups); an American, a New Englander, a resident of Belmont, Massachusetts (another chain); a Jew, a conservative Jew of Eastern European descent (yet another); a musician, a member of the Civic Symphony, a member of the wind section, one of the clarinets (still another)? And in terms of which of these groups do others identify me in this situation? This is crucial because one has to know which of the two axioms to apply.

As with other conceptual categories, a sharp distinction is often presumed between members and nonmembers; there is a demand for some sort of "purity." In academia, for instance: many linguists don't consider me exactly a linguist, because I think about psychology, but many psychologists don't consider me a psychologist, because I don't run experiments. Often groups enforce this purity by establishing systems of admission to the group and procedures for determining descent (including, in academia, intellectual descent). An insistence on group purity combined with the inevitable mixtures of group memberships lies behind the questionable status often accorded to those who make the "mistake" of being of mixed background (think of race and religion here).

In order to ensure the continued cohesion of a group, it is necessary to enforce the expectation that members cooperate. Groups therefore invariably have a code of conduct: a set of normative principles, explicit or implicit, that punish group members who fail to cooperate. Fehr and Fischbacher (2004) and Turillo et al. (2002) show that humans are willing to impose such punishments, even at a cost to themselves. Boyd and Richerson (2005) demonstrate that without such principles a group is inevitably vulnerable to defection. They also show that even the threat of punishment is not sufficient to stabilize a group. It is necessary also that its members have a general cognitive drive to conform to group behavior, in particular for everyone to be willing to punish defectors.

My sense is that the code of conduct is conceptualized as a *joint commitment* of the members ("*We* are committed to these norms"; see section 5.10 and chapter 8). This means that punishment for violating the code, even if carried out by an individual, is conceptualized as collectively imposed by the group or on behalf of the group. One of the worst sanctions that can be imposed, universally I believe, is expulsion from the group—the victim "loses his identity." In smaller, less formal groups, the sanctions may actually be imposed collectively—everyone snubs or retaliates against the offender (Ellickson 1991). Larger, more complex groups have to invent institutions that grant authority to certain individuals to impose sanctions on behalf of the group.

However, even if the "will of the group" is conceptualized as a joint commitment, not everyone need be individually committed to it. The most obvious such case, all too common, is when someone designated as authority defects from the group commitment in order to benefit him- or herself. A more complex case, alas also familiar, is a society in which the authorities fraudulently purport to represent the will of the people, the laws fraudently purport to be for the benefit of the group, and those not in authority publicly play along with the charade out of fear rather than commitment.

5.7.2 Groups as "Superindividuals"

Two criteria seem to characterize the entities I want to regard as groups (e.g. clubs but not the collection of people on the bus):

Criterion 1 The group on occasion acts *as a group* or *in the name of the group*, regardless of whether all the members are involved in the action.
and/or
Criterion 2 The group's existence does not depend on particular people being members; members can come and go but the group remains in existence as "the same group."

A group that meets these criteria has some sort of identity independent of its members. I'd like to consider the hypothesis that when these criteria are met, people tend to conceptualize a group as a *superindividual* and thereby apply the logic of individuals to it. (It is not clear whether non-human primates share this conceptualization.)

Consider some of the hallmarks of group membership. Just as one has one's own self-esteem as an individual, so one has self-esteem that derives from one's group membership—from the "joint self-esteem" of the group. Members have feelings of pride in their own group and a sense of its superiority to other groups. Groups characteristically stage events that reinforce this group identity and allegiance. For instance, rituals that grant membership or status, such as coming-of-age ceremonies, coronations, marriages, and award ceremonies, are not just for the benefit of those who undergo the ritual: they are also for the benefit of the spectators. Other mass events such as funerals and football games also function to strengthen the sense of the group and the concomitant senses of joint commitment and group self-esteem.

Within this superindividual, a group member is conceptualized not as an individual but as an instance of a category, a replaceable "cog" in the larger machinery. On occasion, members may even experience a partial loss of individual ego within the group identity, especially in the context of mass events such as group rituals and wars. Thus, as in all other cases of categorization by humans, there is a pressure to conceptualize all the instances as being alike—to reduce everyone in the group to an essentialized stereotype. (In case intuition doesn't make this abundantly clear, see Hirschfeld 1996 for discussion of racial and ethnic essentialism.) This pressure is not confined to one's conceptualization of other groups: as noted above, within the group there is also pressure for everyone to be alike.

The view of a group as a superindividual also makes it easy to understand the relations among groups. Like an individual, one group can exert dominance over another, compete with another, or form alliances for cooperation with another. In turn, these relations are "inherited" by members of the group. Thus a member of a dominant group will presume personal dominance over a member of a subordinate group (one of the bases of ethnic discrimination and racism). And members of allied groups are more likely to show affiliative behavior than members of competing or hostile groups ("My country is an ally/enemy of your country; therefore you are my friend/enemy").

Gilbert (1989) discusses a tension between Weber's and Durkheim's approaches to sociology. The former insists that society can only be thought of in terms of the sum of individual behaviors; the latter that society is something over and above its members. In a sense, the present hypothesis unites these two views. We are studying the behavior and conceptualization of individuals. But among those conceptualizations is that of the group as a superindividual that transcends its members, and the social behavior of individuals is deeply affected by that conceptualization. (I am not sure that this is Gilbert's own resolution of the issue, though.)

5.7.3 Questions of Learning

In studying other cultures and in engaging in our own, we take for granted all these parts to the logic of groups. This raises the usual developmental issue: Do children learn all this? Or do they have an innate understanding of the logic of groups and just plug into it any groups with which they come to associate?[9] Given that a parallel though less complex instantiation of this logic appears in primate societies, I would be inclined to vote for a substantial innate component. We might also ask whether there are socially impaired individuals who never understand this logic and whether it can be disrupted by brain damage.

What sorts of things might children *learn* about their groups? First of all, of course, they have to learn which groups they belong to, who else is in them, and what other groups are in their social milieu. But there is more to be learned. One important variable in the customs of a group is the degree to which it enforces conformity and sublimation into the group. For instance, it is often said (e.g. Gardner 1983) that American society, at least outwardly, encourages individualism and tolerates nonconformity, whereas Japanese society tends to discourage both. If true, this is an example of a learned difference in group behavior.

Another variable seems to be the intensity with which axiom 2 ("Compete with those who are not group members") is applied to various other

[9] Famous experiments by Sherif and Sherif (1966) showed that children who were arbitrarily divided into two groups spontaneously developed all the stereotypical symptoms of group identification and group competition. This doesn't necessarily show that the logic is innate, given that the children had no doubt experienced plenty of groups. Nevertheless, it does illustrate that powerful forces of social cognition are at work.

groups. For instance, Islam in the Middle Ages and in the Ottoman Empire seems to have had a live-and-let-live attitude toward other religions and ethnicities, in sharp contrast with contemporary fundamentalist Islam. The catastrophe of the early 1990s in the former Yugoslavia can be seen in large part as coming from a radical shift in the official face of this parameter, from relative tolerance to intense intolerance of other ethnic groups (though *un*official intolerance was always rampant). I don't need to multiply the horrible examples. Publicly, especially since the 1960s, American society applauds tolerance. Nevertheless, the impulse toward tolerance is far from universal in the United States, and even those who advocate it often act otherwise. In any event, the settings of this variable (in both its overt and its tacit manifestations) must be learned by individuals from their culture.

A further issue is the dynamics of groups. So far I have talked as though groups last forever. But it is also necessary to understand how groups come to be formed, how they disintegrate, and how they fission. Here is, I suppose, one place where politics enters (another is in challenges to the dominance hierarchy).

5.8 Cooperation and Competition

Let us turn now to more evanescent relations among persons, beginning with cooperation. Cooperation is more than two people acting in a way that coincidentally happens to benefit them both.[10] The key to understanding cooperation comes from the hypothesis (Gilbert 1989; Cohen and Levesque 1991; Searle 1995; Clark 1996; Bratman 1999) that we are capable of conceptualizing "joint actions" and "joint intentions": not just "*I* am doing such-and-such intentionally and *you* are doing such-and-such intentionally," but "*We* are doing such-and-such out of a joint intention, and my role in it is such-and-such and your role in it is such-and-such." A simple case is moving furniture together. My lifting one end of the couch and pushing makes no sense outside the context of your lifting the other end and pulling. But the two together make sense as a jointly intentional action of moving the couch out the door. Another case is playing a duet: it is not just my colleague Valentina playing the piano and me simultane-

[10] The latter is called "cooperation" in the standard prisoner's dilemma scenario, where neither individual has any idea what the other is going to do. This is not cooperation in the sense I'm interested in here.

ously playing the clarinet; rather, it is a jointly intended task incorporating our individual contributions.[11]

The notion of joint intention seems just right to characterize cooperation and agreement. For instance, it is a necessary component of any sort of transaction, trade, or contract. In turn, contracts of course include marriage contracts, which are fundamentally declarations of joint intention to maintain sexual exclusivity (Brown 1991, citing Goodenough 1970). Clark (1996) argues that joint intention is also a basic aspect of linguistic communication. Using language does not just involve a speaker imposing information on a hearer; a conversation is a speaker and a hearer performing a joint task of getting information across.

A joint intention goes beyond standard theory of mind. It does not just conceptualize the minds of others; it sort of pretends that two people are sharing their minds. However, Cohen and Levesque (1991) and Grosz and Sidner (1999) point out that a participant in a joint task does not have to know all the details of the other's action—just enough to ensure that the interaction takes place properly. For instance, I don't know how to play the piano, but I can still play duets with Valentina. The best approximation individuals can actually make to "sharing minds" is for each of them to get assurance from the other that the joint intention is shared. So there have to be signals for offering cooperation, for taking up the offer, and for coordination in the course of the joint task. These can be explicit in things such as stereotyped utterance forms (*Would you care to...?*; *Let's...*; *OK*), signatures on a contract, and a handshake upon reaching an agreement. Bangerter and Clark (2003) trace in some detail how speakers and listeners use expressions like *OK*, *right*, and *uh-huh* to signal that they are in synch.

Alternatively, the coordinating signals may be more subtle, such as little inflections of body language. As suggested in section 4.2, a handshake itself is a joint action. Someone offering a hand is the invitation to joint action; the uptake is the other person's response. Little proprioceptive cues determine the coordination of how hard to grasp, the number of shakes, and when to let go. In playing a duet, coordination is accomplished through close attention to the other's playing, complemented by

[11] This proposal is not without controversy. For instance, Grosz and Sidner (1999) attempt to eliminate the "we intend" from a characterization of joint action; on the other hand, as Hobbs (1999) points out, they replace it with a primitive called "SharedPlan," which has the same effect.

eye contact and physical gestures such as head movements. (See Sebanz, Bekkering, and Knoblich 2006 for some experimental evidence.)

Under what conditions is a joint intention formed and executed? It's perhaps of interest to explore a case on the borderline. You're walking toward someone in the street, and both of you swerve slightly to avoid a collision. This need not require a joint intention, merely a coincidentally coordinated action. But if by chance you both swerve in the same direction at the same time, a funny little dance begins, replete with eye contact and facial expressions, until you manage to find nonintersecting trajectories. This has more of the signs of a joint task.

When chimps and wolves cooperate, do they make use of such cognitive structures? (As suggested in section 5.2, I don't think ants do.) Tomasello et al. (2005) argue that only humans have fully shared intentions and that shared intentions are necessary for both language and culture. Watanabe and Smuts (2004), however, raise a case not considered by Tomasello et al., namely play-fighting, which bears some hallmarks of a joint task and which occurs in various species. The delicate negotiations among primates surrounding who gets to groom who (e.g. Cheney and Seyfarth 1990) also have the feel of joint tasks.

If chimps and wolves can't conceptualize such a thing as a joint intention, what evolutionary precursors *do* they have that enable them to cooperate to the extent they do? One important precursor is surely joint *at*tention: observing that the other individual is looking where you're looking—which doesn't require theory of mind (see Sebanz, Bekkering, and Knoblich 2006 on the role of joint attention; also see chapter 6). Another place to look for precursors might be sexual behavior, both courting and actual intercourse, both of which call for a certain amount of physical coordination between individuals, though not theory of mind ("Birds do it, bees do it, even educated fleas do it").

When an individual is in the course of executing a joint task, prisoner's dilemma sorts of situations arise from the possibility of defection—that is, from the possibility that one participant will abandon a joint intention and leave the other in the lurch. Joint intentions also open opportunities for deception: letting the other participant continue to think one is pursuing the joint project, while actually planning defection at some later point in time. So theory of mind and so-called cheater detection (Cosmides and Tooby 1992)—psyching out whether the other participant *really* shares the joint intention—come into play quite naturally.

On the other hand, sometimes it is quite tolerable to be deceived. Consider mother-infant interaction. The mother interprets the action as joint intention—that's part of what makes her feel bonded. But I doubt the infant is yet capable of seeing it that way!

The notion of joint intention intersects with the notion of a group as a superindividual in an interesting way. As pointed out by Solan (2005), we easily speak of the intention of a legislature, a court, or a nation, even though none of these has a mind of its own. It seems plausible that we conceive of these as intentions of the superindividual, cashed out as "collective intentions" of the group members.

A sort of opposite to joint intention is competition. The participants both know they're out to get each other, so there is in some sense a "joint project" of exploiting each other. This is more complex than just "I'll exploit you and protect myself" (i.e. plain aggression), because it includes a theory of the competitor's goals. But competition is not symmetrical with cooperation: one cannot take advantage of a competitor by defecting. Usually the only way to opt out of competition is to surrender, unless it is by a joint decision (i.e. a cooperative decision) to abandon the activity.

In the previous section, I spoke of the "joint commitment" by the members of a group to the group's code of conduct. A joint commitment is related to an individual commitment in the same way that a joint intention is related to an individual intention: "We are committed to upholding this norm: my role in upholding it is such-and-such, and each other person's role is such-and-such." The possible disparities discussed there between joint commitments and individual commitments correspond to the possibilities for defection from a joint intention. Another related concept is "joint belief": "We believe this proposition: I believe it, and so does everyone else in the group." Section 8.8 returns to joint intention, commitment, and belief, piggybacking on the formal treatment of intention and belief in the earlier parts of chapter 8. Section 10.5 discusses the role of joint intention in exchange transactions.

5.9 Framing

Structures like cooperation, competition, dominance, and group membership have to be integrated dynamically into one's understanding of the situation from moment to moment, to help determine one's course of action. One of the elements of integration might be called *framing* (a term borrowed from Goffman (1974); Minsky (1975) uses the term similarly but with more limited scope).

A typical example of a frame is a concert, in which performers, audience, and ushers have specific roles to play throughout the event—and all participants have to conceptualize it that way in order to understand what is going on and how to behave appropriately in their respective roles. Once the concert is over, the roles are no longer relevant, and the participants drop the frame.

The same event may have different significance depending on the frame in which it is interpreted. One case, mentioned in section 5.5, would be the framing of an act of giving as either tribute, largesse, or expression of affection. For another case, consider a song sung by a character in a movie. If the movie is relatively "realistic," this portrays an actual act of singing in the narrative frame, perhaps a singer rehearsing a love song. But if the movie is a musical, the very same love song may depict a lover breaking into spontaneous song in a way that real people do not. Within this frame, it's even perfectly acceptable for people to hold conversations in song, rhymes and all. (Thanks to Dan Dvorak for this example.)

A more general sort of frame is illustrated by games—which, as far as I know, are found in all cultures. What we see most saliently in a game is the element of competition: two (or more) people knowingly trying to outdo each other. But, as Searle (1995) and Bratman (1999) point out, this competition is set within a larger framework of cooperation: the participants agree to play, and they agree to abide by set rules and a presumption of fairness. If a game were *only* competition, poker players would simply be trying to steal each other's money rather than sitting around civilly at tables. The cardsharp is of course being deceptive about the frame of cooperation (or placing it in a still larger frame of exploitation): he or she *is* trying to steal the others' money.

If someone breaks a rule during a game, the game is suspended till matters are set right. Sometimes the game itself has "metarules" to deal with such cases, but sometimes a violation of the rules instead signifies a defection from the larger frame of cooperation, and the situation degenerates into haggling or even violence.

If the framing of a game is competition within cooperation, does the opposite also exist: cooperation framed inside competition or hostility? This seems a good characterization of *bargaining*: each participant wants things the other has and is trying to get as much as possible while minimizing his or her own losses. But this competition is carried on with a facade of civility, so the participants don't end up knifing each other or stealing outright. (This framing corresponds nicely to Fiske's (1991) "Market Pricing" and Jacobs's (1994) "commercial syndrome"; see section 10.5.) Another example of this framing is the behavior of rival law-

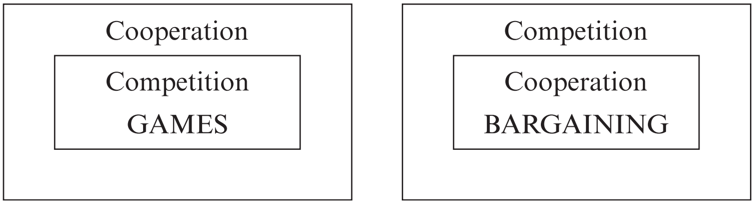


Figure 5.1
Nesting of frames

yers and debaters, who (normally) behave with complete respect for one another, despite their incompatible goals. (Figure 5.1 represents this nesting of frames inside one another.)

Team sports represent a further level of complexity. Each team forms a group whose joint intention is to compete with the other team, the whole within a frame of cooperation with the other team. Think about what this entails in real time. At any moment in a football or hockey game, an individual player has to gauge his or her actions with respect to multiple individuals, establish joint subintentions *very* fast, and coordinate them with physical activity (Horgan and Tienson 2006). A more highly structured game like baseball adds turn-taking to this constant reframing, where the turns are defined both at the level of individual batters and at the level of teams. The whole complex overarching frame of a team game can be adopted on the spur of the moment, as when kids choose sides in a pickup game; or it can constitute a lasting relationship, where teams are affiliated with larger groups (e.g. the Boston Red Sox) and the games are proxy for larger group competition and serve as ritual events that reinforce group identity.[12]

It is possible for participating individuals to differ in their grasp of the details of a frame and its significance. For instance, in a pickup football game some participants will be far more aware of strategic subtlety than

[12] Shore (1996, chap. 3) offers an interesting discussion of many of these points, in particular the many changing layers of framing in play at once, in the context of baseball.

Larry Bacow (pers. comm.) points out that one's loyalty to a team such as the Boston Red Sox has nothing to do with the individuals on the team, since all the players may be replaced in the course of a season or two. This connects to the notion of a group (here, a team) as a superindividual that serves as the locus of loyalty.

others. Similarly, during a synagogue service one sometimes hears whispered discussions about how a particular bit of ritual is to be performed and why, whether doing it the way it is being done “really counts,” and what should be thought of someone who does it that way. At the same time, many participants are largely oblivious to such hairsplitting distinctions.

It is also worth pointing out that our control of frames isn’t airtight. There is often “leakage,” as when competition within a game leaks out of the frame of cooperation and turns to downright hostility or even lasting enmity. The converse also occurs: bargaining and trading, although at bottom competitive, can lead to affiliative bonding.

This is of course all informal and descriptive. But I think it’s a necessary prelude to asking the harder question, the one that properly belongs to cognitive neuroscience: what computational and/or neural mechanisms do we have to posit in order to produce this ability to frame and reframe recursively, and in order to permit the *learning* of this behavior? Team sports are grasped without effort by 10- or 11-year-old children—though not by other primates, as far as we know. (The treatment of “action schemas” in chapter 4 might be a prelude to a more formal understanding of frames.)

5.10 Rules and Other Normative Principles

Going back to games, let’s think about the rules. Rules show up in many different domains. They take a general form something like this:

$$\text{In frame F (or context C), you } \left\{\begin{matrix}\text{should}\\ \text{must}\end{matrix}\right\}\left\{\begin{matrix}\text{do}\\ \text{not do}\end{matrix}\right\} \text{X.}$$

Spelled out a little more explicitly, this might take the following form:

> In frame F (or context C), if you {do/don’t do} X, consequence Y of good/bad value to you will ensue.

Rules apply only to persons. Though you may play with your dog, you don’t play games with rules with your dog. Nor are dogs subject to obligations, laws, or morals—only their owners are.

Different sorts of rules differ in the kinds of consequences they promise or threaten. Some examples:

- In games, the rules define a temporary frame for action within which various rewards or penalties obtain. Breaking the rules incurs a penalty or breaks out of the frame.

- An obligation (or contract, including a promise) specifies certain actions that the holder of the obligation is to perform for the benefit of the person to whom the obligation is made. If I fail to meet my obligation to you, you get the right to perform some action that harms me. For instance, if I fail to pay off a debt to you, you have the right to demand restitution and perhaps further sanctions against me. Depending on the sort of obligation, you may be entitled to punish me yourself, or you may have to appeal to the group as a whole or to the group's designated authority to impose punishment on me. (Chapter 11 discusses obligations in much more detail.)
- A legal code designates certain actions as desired or sanctioned by the authority of the group (whether or not assented to by group members); the consequences of reward or punishment are carried out by designated representatives who act as proxy for the group. This is "institutionalized morality," to use the term suggested by Alexander (1987).
- A system of moral or ethical rules designates certain courses of action as morally good and others as morally bad (and leaves the rest neutral). As far as I can see, the consequences associated with moral rules generally concern the approval and trust of community members. If you do something morally good, people think more of you and trust you more, and if you do something morally bad, the opposite. Trust in an individual, in turn, translates into increased interest in cooperative enterprises with this individual (Trivers 1971; Alexander 1987; Ellickson 1991; Fehr and Fischbacher 2004; see also chapter 9).
- Religious codes replace approval by the community with approval by deities or other supernatural beings such as ancestors (Fiske 1991; Boyer 2001; Atran 2004). In the Judeo-Christian tradition, the consequence isn't just approval or disapproval, it's specific reward or punishment, perhaps in the afterlife. Jewish tradition even sees its religious codes as a legal contract between God and the group.

One could go on and cite many other kinds of rules: parents' rules for their children, manners and rules of etiquette, dress codes, dietary customs, and so on. I think, though, that they are basically all of the same form; they differ only in the frames within which they are applied and in the general form of the consequences. Each type of rule attaches a social value to a kind of action—again connecting the social plane to the physical. A group's code of conduct is made up of such rules, explicit or tacit.

The language used to express all these sorts of rules is also pretty much the same, involving the use of the modal verbs *should*, *must*, and *may* and adjectives such as *right* and *wrong*.[13] As a consequence, they are not always clearly distinguished. For instance, moral/ethical codes are often taken to be based entirely in religious codes. To be sure, aspects of religious codes often do state moral principles. But that does not make them the same. There are ethical codes that are independent of religion, such as honor among thieves and perhaps desert traditions of hospitality. And many religious codes such as principles for performing rituals hardly fall in the moral domain.

Similarly, some writers on promises (see e.g. Conison 1997) conflate the contractual and ethical domains. For them, the important thing about promises is that it's morally bad to break them. They treat the consequence of breaking a promise as disapproval by the community, that is, as an ethical breach. This account neglects the fact that there is simultaneously a contractual breach, which gives the individual to whom the promise is made very specific rights. Thus breaking a promise has a consequence in both domains. Similarly, legal contracts (as opposed to mere private agreements between individuals) have consequences in both the contractual and legal domains. (See chapter 11 for more discussion.)

A particular action may have conflicting consequences in different normative domains. A classic case is the evil landlord in the melodrama, who is foreclosing on the poor widow in exercise of his contractual right, but in so doing violates the moral code. Conversely, nonviolent civil disobedience along the lines of Gandhi and Martin Luther King violates the legal code but conforms with what is taken to be a higher moral value. Sripada and Stich (forthcoming) discuss how norms can be independent of social institutions and laws. More generally, my sense is that at best, explicit legal and religious codes are intended as codifications of a more inchoate sense of morality; Mikhail (forthcoming) stresses the way issues involving legal codes mirror intuitive judgments of morality. On the other hand, legal and religious codes can be used to legitimate the raw exercise of

[13] These are not the only uses of these words, though. They also have a "prudential" sense, used for giving advice: *It's raining, so you should take an umbrella*. This sense can be distinguished from the "rule" sense by noting who benefits: *You should take an umbrella = It would be good* ***for you*** *to take an umbrella* versus *You should write your aunt a thank-you note = It would be good* ***of you*** *to write your aunt a thank-you note*. The former is for *your* benefit and is therefore prudential; the latter is for your aunt's benefit and therefore normative. The modal verbs also have a predictive sense: *The bus should arrive soon*; *The bus must be there by now*. See sections 9.3 and 9.6 for more detail on these distinctions.

power; Alexander (1987) (like many others) observes that all too often laws are made primarily for the benefit of lawmakers and their social circle.

The distinction among these rule types goes further. Widely cited experiments by Turiel (1983) show that young children readily distinguish those norms that we could "decide to change" (the social conventions) from those that we could not (genuine morality). The latter are taken to be timeless, universal, and objective. Following this lead, as well as the lead of standard moral philosophy, a recent strand of research in cognitive neuroscience (e.g. Mikhail, forthcoming, and possibly Hauser 2006) has taken the view that morality can and should be studied in isolation: in order to get at the root of human nature, we ought to strip away the relative superficiality of social convention, looking for universal principles of moral judgment.

I disagree with this stance for four reasons. First, cultures differ in what they themselves consider to be morality as opposed to social convention, particularly with respect to issues such as sexual mores and slavery. Two hundred years ago, there were large portions of the world where slavery was considered morally acceptable. Does that make the status of slavery just an issue of social convention? We wouldn't say so now. Even within a culture there may be differences among subcultures, each of which regards its own sense of morality as the only proper one. Lakoff's (2002) "strict father" versus "nurturant parent" models of morality might be such an example; Doris and Stich (2005) cite issues such as abortion, capital punishment, and gay marriage as cases where within American culture people's moral judgments diverge radically.

A second reason to avoid the temptation to look for universal morality is that when we look at a culture from the outside, what looks to us like social convention and what looks to us like morality are inextricably intertwined. Consider for instance the Ten Commandments, where alongside the moral dictate "Thou shalt not murder" is what looks like a social convention: "Keep the Sabbath." Nevertheless, the book of Exodus makes no distinction: violating either one is punishable by death (21:12 for murder, 31:15 for working on the Sabbath).

A third reason why morality cannot be universalized is that it is deeply tied up with the logic of groups and framing. A group may recognize its own particular variations on morality: "We hold ourselves to a higher standard" or "If we do this, our gods curse us, but it's different for you." In the frame of a war, killing is typically regarded as good rather than immoral, if the victim is the enemy. And an advocate of capital punishment believes that killing as punishment for certain crimes is morally acceptable or even good.

More generally, the sociologist Jane Jacobs (1994) proposes that human societies universally have two independent systems of morality, with partly contradictory tenets. One, the "guardian syndrome," concerns norms for keeping the group cohesive and defending it against aggression from other groups. The other, the "commercial syndrome," concerns norms for participating in trade with other groups. The two moral systems coexist, if uneasily, in every successful society; the trick is to know when each is appropriate (an issue of framing).

But there is a fourth reason—a theoretical reason—not to isolate morality from social convention. As intimated above, one of the motives for attempting this separation is to discover "genuine human nature" beneath the variations of culture. But such a view incorrectly takes the variations of culture to be relatively superficial and uninteresting. Recall the analogy to the study of language (section 5.2). Universal Grammar is not a theory of what is universal in language; it is a theory of children's ability to *learn* language. It has to take into account the range of variation among languages, what happens frequently and what never happens, what's easy for children to learn and what's difficult. Similarly, in seeking to discover the aspects of human nature underlying human society, we cannot just insist on the universals of culture: we should be looking at the range of variation and how the common issues of humanity play out crossculturally. Thus the system of norms as a whole seems a more ecologically appropriate object of study.

I find it intriguing that normative rules—of all sorts—are taken to be objective entities in the world, albeit abstract. We face a certain cognitive dissonance. On one hand, we know that people made them up. But on the other hand, they're hardly imaginary! Within a game, I objectively win or lose. If I break a promise, or if I fail to pay my taxes, the consequences are real. So rules, once they are established as consensual, are practically as irresistible as laws of physical causality.

In particular, as mentioned above, rules that we call "moral" are conceptualized as timeless, universal, and objective, whether or not they really are crossculturally and historically. This is why moral relativism is so repugnant to many people: they reason that if a rule is relative, it *can't* be moral. (Chapters 7 and 9 address this issue.) Doris and Stich (2005) point out that this folk stance on morality is taken over as the basic hypothesis of a major school of moral philosophy, moral realism. They offer numerous counterexamples to this premise and propose a cognitive approach to morality along lines similar to the present discussion.

Again, it would be of interest to explore the brain basis and the developmental course of rules. How do rules differ from pure principles of associated stimulus and response? Do children understand rules in the same way in all normative domains? How are rules learned? At what age can children learn games? Most of the rule systems I've mentioned purport some sort of impartiality or fairness. What do children of various ages think fairness is? Are there brain deficits that lead to failure to understand rules? And so forth. The work of Piaget (1932), Kohlberg (1981–84), and Turiel (1983) is at least a starting point.[14]

To be sure, the codes of conduct particular to a community have to be learned by children, as well as by outsiders who interact with or join the community. But it's quite possible that we don't have to learn *that there is such a thing as a code of conduct*. Rather, the pervasiveness of such organization suggests it is a skeletal conceptual structure around which humans organize their social existence. We return to this issue in chapters 9 and 11.

5.11 What Grounds Morality? Where Science Bumps Up against Politics

Many of the issues touched on here are very tricky—and not just scientifically tricky. They run below the surface of a lot of intense public debate, not to mention thousands of years of philosophical, religious, and political discourse. The underlying question is, what are the sources of principles of fairness and of moral/ethical values, particularly those that are conceptualized as universal and timeless? For instance, what justifies the stance in the United States that favors tolerance of other ethnic and religious groups? Why shouldn't we instead applaud efforts to teach xenophobia and white male supremacy?

A great deal of Western and especially American tradition has regarded moral values as given by God, for instance in Jefferson's phrase

[14] For a trenchant critique of the Kohlberg "stages," see Macnamara 1991. The Very Brief Version is that if young children had the view of morality attributed to them by Kohlberg, they would not be able to understand the point of fairy tales like "Cinderella." Kohlberg, following Piaget, claims that children believe that one should behave a certain way just to avoid punishment (i.e. the *should* is understood as prudential rather than moral). Were this the case, they could not feel moral outrage at Cinderella's treatment by the wicked stepmother.

"endowed by their Creator with certain unalienable rights" (though there is some question whether such deistic grounding was actually intended (Allen 2005)). I am given to understand that Islam takes a similar stance on the grounding of morals. The moment that Darwin's *Origin of Species* was published in 1859, the threat to this position from evolutionary theory was sensed by all participants in the debate, and certainly in the United States this threat is connected with the rise of religious fundamentalism and its continuing hostility to evolutionary theory.[15] For if morals are not given absolutely by God, where do they come from? If morals are relative or subjective, just made up by people, who says you can't make them up any way you want? How can you argue against Nazism or Communism—or drugs or free love? One reads letters to the editor that proclaim secular humanism as the greatest threat to civilization; it's far better to trust in the truth as revealed by God. The consequences of this attitude for education and for public discourse in science and the humanities are obvious. Worse, we still see military actions being justified (by both sides) on grounds of timeless absolute God-given morality.

To my knowledge, no one has offered a coherent answer to the question of how moral values are to be grounded within a society that does not rely on a particular God's authority—that is, within the global society we all live in now. Rawls (1971), for example, proposes the doctrine of "justice as fairness," arguing that this is what people *would* want to adopt, given the proper circumstances. But that is not the same as being able to say what they *should* adopt. Discussing several modern positions that have concluded that the only possible foundation for law is the threat of force, Mahlmann (2003) shows why they are unsatisfactory, and not just because of their unappetizing conclusion. For the most part, I don't find that people opposing the religious fundamentalists and the economic Darwinists really try to answer the question; they just assert their own moral codes and point out the contradictions and vast helpings of self-interest in the religion-based position.

[15] Barzun (1958) recorded that hostility to Darwinism on religious grounds abated during the first half of the twentieth century, as deficiencies in evolutionary theory became more evident. In particular, he took the absence of an explanatory mechanism for inheritance as a defect that palliated the threat of evolution in the eyes of the broader community. He was, of course, writing prior to the discovery of the structure of DNA and the ensuing explosion in understanding of genetics (the first edition was in 1941). One might therefore conjecture that the renewed opposition to Darwinism in recent decades is a consequence of its increased success.

One folk theory, taken up by philosophers such as Locke, Rousseau, and Kant, has it that at some point people sat down and agreed on the "social contract"; this does explain the sense of social codes as joint commitments. Rawls (1971) begins with this premise, while making clear that he intends it as a fiction—he proposes that we should think about a code of justice from the position "as if" we were hypothetically devising a social contract. Now formal legal systems are indeed developed by people sitting down and making them up, but I doubt this is the case with most elements of codes of conduct in most societies. Of course, the parallel (and robust) folk theory of *language*—that people sat down and decided how to say things—is totally implausible. This ought to give us pause for the social case.

I am not so sure that a theory of social cognition can provide a proper grounding for values either, although perhaps it can offer some insight into sources of difficulty. Let me offer two contrasting examples of what I have in mind. First: It's been well established by evolutionary psychologists such as Dawkins (1989) that there is an asymmetry between males and females in reproductive strategy. Reproduction is a small investment for a male: he just has to perform the act. But it's a large investment for females, who have to produce large eggs and (in the case of mammals) nourish the babies. This asymmetry drives lots of behavioral asymmetries observed in lots of species. One particular game-theoretic consequence is that males are more likely (or more inclined) to be sexually promiscuous than females, a phenomenon we observe in humans as well. But we wouldn't want to argue from this biologically driven logic that this is the way it *should* be—that we *should* condone or even encourage male promiscuity. Morality ought to be properly distanced from biology here (notice that I can't evade a normative conclusion: I have to say *ought*). It is the distance between them that creates a constant tension.

Dawkins's way of putting this conclusion is that our rationality can free us from the dictates of our genes. But turning to rationality or science to tell us which course we ought to follow implicitly assumes some particular goal for how we want society to be. And what justifies *that* goal? We are back in the same boat.

A second case: One aspect of the logic of groups is that a group's code of conduct creates pressure from the group to dampen aggression among group members. In effect, the group protects its members from harm by other group members. Thus in all cultures, the legal system (written or unwritten) punishes not only physical aggression like assault, but also economic aggression like stealing. In fact, de Waal (1996) has observed that similar things happen in chimpanzee groups: dominant individuals often step in to break up fights and take the side of the underdog.

Now it's a tenet of modern free market capitalism that businesses *should* be free to exploit people economically and that the government *shouldn't* be allowed to protect citizens from such exploitation (this "distorts the market"). Capitalism is of course based on bargaining as its basic form of interaction, which on the analysis in section 5.9 emerged as a kind of tamed aggression—but aggression nonetheless. Free market capitalism, then, claims that such aggression should be exempt from government regulation (i.e. group constraints); globalization is an attempt on the part of corporations to evade such group constraints, as it were operating as lawless pirates. Thus we might conclude that the basic premises of free market capitalism flout the universal logic of groups, established by our innate capacity of social cognition. In Jackendoff 1994, chap. 15, I took this as an argument against free market capitalism and globalization. But it is an argument only if we accept that in some sense evolution has "made the right choice"—which begs the question (not to mention in the previous case exonerating male promiscuity). The underlying problem is that although evolution has given humans the ability (and the need) to make moral choices, evolution itself does not make moral choices.

Another possible attack on free markets might address the legal fiction that a corporation is a person: this premise is what gives rise to the idea that corporations have rights at all. At a crude level this idea is appealing, growing out of the conceptualization of a group as a superindividual. On the other hand, a corporation does not so clearly fulfill the cognitive criteria for a group. But again, how does one argue how corporations *should* or should *not* be treated, other than in terms of the moral intuitions concerning the effects?

A deeper issue emerges from the basic way we conceptualize ourselves as humans, as a combination of the physical and personal planes. Our bodies and our animacy (the "brute instincts") are encoded in the physical plane. But the parts of us that we hold most precious are encoded in the personal plane: our personal identity, our sense of free will, and our moral responsibility. Now consider: the goal of the Enlightenment, broadly stated, was to discover what we humans are and what our place is in the world, using rational techniques rather than religiously dictated faith. Reason was going to tell us the point of our lives. But beginning with Darwin and especially in the last half of the twentieth century, reason has led us to the conclusion that—there is no point to our existence! Our bodies are the product of an unimaginably long, mindless process of environmentally shaped evolution, and our minds are the product of the activity of an unimaginably large collection of mindless neurons. The soul is a confabulation, our intuitive sense of free will a useful illusion. This conclusion is trumpeted today in a flood of popular books by cognitive neuroscientists (e.g. Crick 1994; Damasio 1994; Pinker 1997; Ainslie 2001; Wegner 2002; and practically everything by Daniel Dennett).

Awe-inspiring though this result may be to a scientist, it is cold comfort to ordinary human beings, who, *because of their mental constitution, cannot help but understand their deepest hopes and aspirations in terms of the personal plane*. The consequence is that when cognitive neuroscientists and evolutionary biologists attempt to educate the public, they are taken as attacking personhood, human dignity, and moral responsibility; the natural reaction is fear and defensiveness. It is not enough for scientists to say "Sure, these results are counterintuitive, but so are relativity theory and quantum mechanics, so you should get used to it." Relativity and quantum theory don't threaten one's personhood. Thus it should be no surprise that one outcome of the scientific view of human beings is a widespread suspicion of science, at the time of this writing extending to the highest levels of government in the United States. The associated turn toward religious fundamentalism is not just about moral values: whatever their other faults, religions grant human beings a central place in the workings of the universe, which is where, by our nature, we deeply want to be. In a sense the Enlightenment, by undermining its original goals, has failed us. (Similar points are made by Alexander (1987, 31).)

I have no prescriptions for how the field ought to deal with these issues. The point of these examples is not to offer a solution, but only to show how a theory of social cognition affects these issues and perhaps sheds some light on why they arise. More generally, I don't think that a theory of social cognition can offer a full grounding for values; at best, it can help us appreciate a fuller range of possibilities (this is Alexander's (1987) and Mahlmann's (2003) conclusion as well). But it's important to remember that these political issues *are* part of the territory. Those of us who want to work in this area ought to be prepared to discuss the questions openly and thoughtfully, bringing to bear our (hopefully) growing understanding of the sorts of cognitive entities moral codes are, of the role moral codes play in the functioning of a society, and of the innate underpinnings of social understanding that help shape moral codes in every culture.

The point is that we never can be just innocent, objective scholars. We have to be alert to potential political consequences of our research. In particular, we should be concerned that our work is not taken up by demagogues eager to make pernicious political points (as happened with both Darwinism and sociobiology).

All right. This chapter has been an extended meditation on big issues for a field of inquiry whose parameters are just beginning to fall into place. Many of the issues I've talked about have been discussed by everyone from the Greeks to the great religious thinkers, and by long traditions in social and political philosophy. What is different in the approach taken here is that we have contemporary tools of cognitive neuroscience at our disposal, which I think in the end can provide a far more comprehensive view of human nature.

PART II

The Structure of Social Cognition and Theory of Mind

Chapter 6

Perception Verbs and Theory of Mind

Part I of this book was devoted to sketching out broad domains of mental structure. We now turn to more detailed investigation of particular families of concepts involved in theory of mind and social cognition.

This chapter deals with perception verbs such as *look* and *see*. We will discover that within the class of perception verbs there is a divide between those that imply a theory of mind and those that do not. One subclass, which includes *see*, treats perception in terms of the experience of the perceiver. The other subclass, which includes *look*, treats perception in terms of observable exploration of the environment, with no implication about the observer's mind. However, there is a lot of fluidity between the two classes, with many verbs such as *feel* serving in both. The moral of the analysis will be that it is likely impossible to isolate theory of mind as a distinct module in the sense of Fodor 1983, or even in the less strict sense of Jackendoff 2002a. Rather, theory of mind consists of a collection of predicates deeply integrated into the system of conceptual structure.

A second issue addressed in this chapter is probably of most interest to linguists: how psychological predicates map to syntactic structure. A longstanding puzzle is why, although the pairs of sentences in (1) and (2) are nearly synonymous, their subjects and objects are reversed.

(1) a. John fears sincerity.
    b. Sincerity frightens John.

(2) a. John regards sincerity as dangerous.
    b. Sincerity strikes John as dangerous.

As pointed out by Carter (1976), such pairs do not exist in most of the vocabulary. For instance, there are no verbs like **benter*, **shmeat*, or **krill*, forming synonymous pairs like those in (3)–(5); and this seems crosslinguistically to be the case.

(3) a. John entered the room. =
b. The room *bentered John.

(4) a. John is eating the apple. =
b. The apple is *shmeating John.

(5) a. Fred killed the fish. =
b. The fish *krilled Fred.

What is it about the semantics of psychological predicates that makes pairs such as (1)–(2) possible? The answer emerges first in the context of the perception verbs; in chapter 7, I will extend it to the treatment of evaluative predicates.

6.1 Introduction to Part II: Overview of Conceptual Structure

As background for the analysis here and in the chapters to follow, a brief overview of my theoretical assumptions and technical machinery is in order.

6.1.1 What Is Conceptual Structure?

My basic premise is that linguistic semantics is to be conceived as part of a larger psychological theory of how humans understand the world, and that the object of investigation is a form of mental structure called *conceptual structure*. As mentioned in chapters 1 and 2, this approach to meaning contrasts with standard philosophical approaches in the Frege-Tarski tradition, in that I do not take conceptual structure to map directly into the real world. Rather, conceptual structure encodes the world *as human beings conceptualize it*, quite a different notion (Jackendoff 1983, 2002a). Conceptual structure is indeed connected and constrained by the external world—but indirectly, via the complex mappings between sensation and cognition that are established by the perceptual systems of the brain. Thus I am trying to study human concepts, not "ultimate reality."

The analysis will be couched in terms of the theory of Conceptual Semantics (Jackendoff 1983, 1990, 2002a). The basic tenets of this theory are these:

- Conceptual structure, which encodes the meanings of words, phrases, and sentences, is a level of mental structure independent from syntax and phonology, and potentially present (to some degree) in nonlinguistic organisms such as apes and babies.

- Conceptual structures are built combinatorially out of elements describable in terms of a formal generative system.
- The level of conceptual structure is one of the loci of thought or reasoning; that is, rules of inference and heuristics can be formally defined over conceptual structures.
- The level of conceptual structure is linked to linguistic structures by *interface rules*, rules that relate distinct levels of representation. Among the interface rules are *words*, which connect pieces of conceptual structure to pieces of syntactic and phonological structure. There are also interface rules that deal with phrase- and sentence-sized structures.
- Conceptual structure can be linked to mental structures involved in perception and action, again by means of interfaces that relate disparate levels. It is these interfaces that permit us to talk about what we see and to translate verbal instructions into actions.

In short, conceptual structure is a level of mental structure that is largely *autonomous* of language and *epistemologically prior* to it. The function of language in the ecology of the mind is to express conceptual structures overtly for purposes of communication. Language also serves, through the medium of verbal imagery, as a means of making thought consciously accessible (chapter 3; Jackendoff 1987; 1997a, chap. 8).

Let me summarize very briefly the differences between this approach and some other approaches to semantics in the literature (details in Jackendoff 1983, 1987, 2002a):

- Chomsky (1995, 2002) speaks of a "conceptual-intentional interface" in language, which is taken to be a representation of meaning. However, because his conception of the language faculty is syntactocentric (chapter 2), all the combinatorial properties of meaning are taken to arise through the syntactic component of grammar. Moreover, for reasons that I have never been able to determine, the study of thought or meaning as an independent generative system is not undertaken within this tradition and is indeed frowned upon.
- Fodor's conception of the "language of thought" (1975, 1983, 1998) grants meaning an independent status, but insists (a) that word meanings have no internal structure, thus discouraging the study of lexical semantics, and (b) that the language of thought is intentional, in the sense of being directly related to the real world, thus discouraging

the study of the relation of meaning to perception (see also remarks in section 1.2).[1]

- Formal Semantics (Chierchia and McConnell-Ginet 1990; Larson and Segal 1995; Heim and Kratzer 1998) correctly treats meaning as the product of a formal generative system independent of linguistic expression, over which rules of inference can be defined. But this tradition usually insists that semantics has little or nothing to do with psychology, and it invests its energies in set-theoretic treatments of reference that do not lend themselves at all to psychological interpretation.
- Cognitive Grammar (Lakoff 1987; Langacker 1987; Talmy 2000) takes seriously the idea that meaning is in the head, but to my mind it is not rigorous enough in its formalization (indeed, many of its practitioners are explicitly antiformal); nor is it very concerned to integrate its results with the rest of psychology. It is moreover skeptical about the need for an independent notion of syntax in the language capacity.
- Finally, Semantic Network Theory (Collins and Quillian 1969; Simmons 1973; Kintsch 1974) has led to interesting experimental results on the psychological properties of individual words and the concepts they express, but it is virtually silent on the issue of combinatoriality: how word meanings are combined into phrase and sentence meanings.

Like any theory of meaning, the theory of conceptual structure should be supported by linguistic (including crosslinguistic) evidence and by its ability to formally support reasoning. However, because it is supposed to be embedded in a larger psychological theory, it should also interact with evidence from perception, child development, and neuroscience. And since conceptual structure is meant to a degree to be independent of the language capacity per se, we should in principle be able to test the theory against evidence from the cognition of animals, especially primates, both in the laboratory and in natural settings.

The domain of concepts investigated most intensively by myself and many others (see e.g. Talmy 1983; Herskovits 1986; Vandeloise 1986; Levin and Rappaport Hovav 1995; Bloom et al. 1996; van der Zee and Slack 2003; Coventry and Garrod 2004) is *spatial cognition*: the position

[1] Despite proclaiming this position for 30 years, Fodor has not produced any work that offers solutions to the problems addressed by serious decompositional theories of word meaning, including the analyses in this and the succeeding five chapters. See Jackendoff 2002a,b on Fodor's treatment of intentionality.

and movement of physical objects and substances in space, the forces they exert on each other, and the temporal structure of the states and events that result. This domain is especially fruitful because there is a vast range of lexical items expressing spatial concepts, and because these correspond to a rich and precise set of perceptually based intuitions. There is now a flourishing investigation into the crosslinguistic expression of these concepts (Bowerman 1996; Levinson 2003), their mapping into syntactic structure (Pinker 1989; Levin and Rappaport Hovav 1991, 1995), and their developmental course both linguistically (Bowerman 1996; Landau 1996) and nonlinguistically (Carey 1985; Baillargeon 1986; Spelke 2003). In addition, it has long been recognized that language expressing spatial concepts is mirrored to a considerable extent by language expressing concepts in other domains (among many others, Gruber 1965; Jackendoff 1976; Lakoff and Johnson 1980). Consequently, understanding the organization of spatial concepts helps set a foundation for investigating other domains.

However, in this profusion of work, little progress has been made in analyzing "psychological predicates" that involve an Experiencer's state of mind, such as *believe*, *intend*, *see*, and *happy*. And many important social predicates such as values and obligations have never been seriously addressed. So the analyses in this chapter and the next five venture out into new territory.

The analysis inevitably involves a degree of formalization that, while not to every reader's taste, enables me to express various generalizations more compactly and to make the claims of the analysis clearer. Throughout the exposition, I will explicate the formalism in ordinary language as much as possible. However, it is worth attempting to follow the formalism, because it predicts the existence of patterns that we would not otherwise look for, and it reveals significant generalizations and insights that less formal methods do not make evident.

6.1.2 Thematic Roles

Let me give a little overall flavor for the formal analysis. An important part of the compositionality of meaning is captured by treating a situation in terms of the roles its characters are playing. These are delineated in terms of *thematic roles*, which in turn are identified as particular argument positions in conceptual functions. Following Gruber's (1965) analysis, the central role in the system is *theme*, the character whose location, motion, or change is being asserted; this is the first argument of the functions BE (for location) and GO (for motion or change). The role *goal* is the argument of a function TO, which in turn serves as a possible realization of the second argument of GO. In a typical motion sentence like (6), then, the subject is theme and the object of the preposition is goal; the prepositional phrase as a whole constitutes a *path.*[2]

(6) Phonology/syntax: The ball rolled to the wall.
Conceptual structure: BALL GO(+manner) [TO WALL]
theme goal (arg of TO)
|
(1st arg of GO) path (2nd arg of GO)

Another important thematic role, usually called *agent*, is the first argument of the function CAUSE. It should be noted that agency does not require animacy or volition. As seen in (7), the wind can be an agent in this sense. The second argument of CAUSE is an event that might be called the *effect*.

(7) Phonology/syntax: The wind made Bill sneeze.
Conceptual structure: WIND CAUSE [BILL SNEEZE]
agent effect

It is important to understand that the markings "theme," "goal," and "agent" are merely notational heuristics and play no part in the formal analysis. Thus when we speak of a verb "assigning thematic roles" to its

[2] In an effort to make the notation a bit more reader-friendly than that in Jackendoff 1990, I am adopting the following equivalents:

[BE (X, Y)]	is replaced with X BE Y
[GO (X, Y)]	is replaced with X GO Y
[CAUSE (X, [...])]	is replaced with X CAUSE [...]
AFF (X, Y)	is replaced with X AFF Y
AFF (X,)	is replaced with X AFF
AFF (, Y)	is replaced with AFF Y

The result conforms more or less to general practice (e.g. Levin and Rappaport Hovav 1995). For a tree notation that also has its advantages, but uses a lot of space, see Jackendoff 2002a.

A further point: Standard convention would simply call the first line of (6) "syntax." Here I am somewhat compulsively insisting on the approach of Jackendoff 2002a and chapter 2, in which the pronunciation belongs exclusively to phonology and does not appear at all in syntax. Syntax contains *only* syntactic features such as part of speech, number, grammatical gender, and case.

subject and object, what we really mean is that it instantiates the arguments of its meaning with the meanings of its subject and object.[3]

A verb need not express a simplex function; it can "incorporate" two or more functions into its meaning. For example, *enter* means essentially 'go into'; and the transitive verb *roll* incorporates the meaning of the intransitive *roll* shown in (6), so it means 'cause to roll'. The parts of conceptual structure corresponding to the verbs' meanings are underlined in (8).

```
(8) a. Bill    entered       the room.
       BILL GO [TO [IN ROOM]]
       theme                goal
    b. The wind rolled   the ball                to the wall.
       WIND    CAUSE [BALL GO(+manner) [TO WALL]]
        agent             theme                       goal
```

6.2 Actors and Patients/Undergoers

Chapter 2 proposed that conceptual structure is parceled into *tiers*, semi-independent structures that reflect different aspects of meaning and conceptualization. These included the *propositional tier*, which encodes who did what to whom, and the *information structure tier*, which encodes the division of sentence meaning into topic, focus, and common ground (or old and new information). Elsewhere, I have also proposed a *referential tier*, which encodes claims of referentiality and scope of quantification (Jackendoff 2002a, chap. 12). Within the propositional tier, I have proposed a further distinction between the *thematic tier*, the aspect of meaning just discussed, and a further layer of functions, the *action tier*, which lays out a somewhat cruder overview of the action in a sentence (Jackendoff 1990).

Section 6.3 will extend the action tier to some of the perception verbs, which are not action verbs. Since the term "action tier" thereby becomes too narrow, I will adopt a term used by Van Valin and LaPolla (1997) and call it the *macrorole tier*.

[3] I stress this because mainstream syntactic theory often speaks of "theta-marking" (where "theta" abbreviates "thematic"), as though it consists of assigning simplex role-names like "theme" to NPs in syntactic structure—quite a different notion, and one far less directly related to semantics. It is not possible to define inference rules over a syntactic tree annotated with theta-role names. See Jackendoff 1990, sec. 2.2, for more discussion of the differences.

The motivation for the macrorole tier comes from considering the roles Actor—the character doing the action—and Patient—the character affected by the action. The standard test for an Actor is the context *What X did was*. . . . For instance, in (9a–c) the subject is an Actor, but in (9d–f) it is not.

(9) a. The ball rolled to the wall. What the ball did was roll to the wall.
b. The wind made Bill sneeze. What the wind did was make Bill sneeze.
c. Bill entered the room. What Bill did was enter the room.
d. The ball was in the corner. *What the ball did was be in the corner.
e. The wall surrounded an orchard. *What the wall did was surround an orchard.
f. Bill owned a VW. *What Bill did was own a VW.

The standard test for Patient is *What happened to X was*. . . . In (10a,b), the direct object is Patient; in (10c,d), it is not.

(10) a. Bill ate the apple. What happened to the apple was Bill ate it.
b. The wind knocked Bill over. What happened to Bill was the wind knocked him over.
c. Bill liked the apple. *What happened to the apple was Bill liked it.
d. The wall surrounded an orchard. *What happened to the orchard was the wall surrounded it.

The surprise is that the roles Actor and Patient are to some degree independent of the standard thematic roles theme, agent, goal, and so forth. For instance, in (11a), *Bill* is the initiator of action and therefore agent; *the ball* is in motion and therefore theme. At the same time, *Bill* is Actor, as seen from (11b); and *the ball* is Patient, as seen from (11c). The combination of roles is thus as shown in (11d).

(11) a. Bill threw the ball.
agent theme
b. What Bill did was throw the ball.
c. What happened to the ball was Bill threw it.
d. Actor = agent; Patient = theme

(12a) presents a different configuration. *The car* is in motion and therefore theme; *the tree* is the endpoint of motion and therefore goal. The tests for Actor and Patient reveal two different construals: (12b–d) and (12e,f). (12d) pairs Actor with theme, whereas (11d) and (12f) pair Patient with theme. (12) also illustrates that Actors need not be animate, much less volitional.

(12) a. The car hit the tree.
theme goal
Construal 1
b. What the car did was hit the tree.
c. What happened to the tree was the car hit it.
d. Actor = theme; Patient = goal
Construal 2
e. What happened to the car was it hit the tree.
f. Patient = theme

For another example, consider (13a,b). In both cases, the books are moving onto the truck, so *the books* is theme and *the truck* is goal.

(13) a. Bill loaded the books onto the truck.
b. Bill loaded the truck with the books.

But the two differ in how naturally the two phrases can be construed as Patient.

(14) a. *(From (13a))*
What happened to the books is Bill loaded them onto the truck.
?What happened to the truck is Bill loaded the books onto it.
b. *(From (13b))*
?What happened to the books is Bill loaded the truck with them.
What happened to the truck is Bill loaded it with the books.

This corresponds to the standard intuition that the direct object is the entity "affected" by the action.[4] However, this is not a necessary property of direct object position. For instance, the objects in the following sentences are not Patients, even though in (15a,b) the subject is an Actor.

(15) a. Bill entered the room.
cf. *What happened to the room was Bill entered it.

[4] For an account of why (13b) is completive (the truck ends up loaded) but (13a) need not be, see Jackendoff 1996a, which includes discussion of other popular proposals such as the one offered by Tenny (1994). In particular, examples like those in (15) make it impossible to identify direct object position with "affectedness," a central aspect of Tenny's position.

b. Andy uttered the answer.
cf. *What happened to the answer was Andy uttered it.
c. The doctor underwent an operation.
cf. *What happened to the operation was the doctor underwent it.

So far all the examples have involved transitive verbs. Looking at intransitives, we see that many favor Actor subjects (16a) and a few favor Patient subjects (16b).[5]

(16) a. Bill strutted/jogged out of the room.
cf. What Bill did was strut/jog out of the room. (Actor)
*What happened to Bill was he strutted/jogged out of the room. (*Patient)
b. Bill died/got sick.
cf. *What Bill did was die/get sick. (*Actor)
What happened to Bill was he died/got sick. (Patient)

But many intransitive verbs whose subjects are theme are indifferent as to whether their subjects are Actors or Patients, the latter construal emerging more prominently when the subject is inanimate.[6]

(17) a. The ball rolled down the hill.
cf. What the ball did was roll down the hill. (Actor)
What happened to the ball was it rolled down the hill. (Patient)
b. The chocolate melted.
cf. What the chocolate did was melt. (Actor)
What happened to the chocolate was it melted. (Patient)
c. The car broke down.
cf. What the car did was break down. (Actor)
What happened to the car was it broke down. (Patient)

On the basis of these sorts of observations, in Jackendoff 1990 I proposed a conceptual function $X\ AFF^{-}\ Y$, roughly 'X acts on/affects

[5] The term "Undergoer" is sometimes used instead of Patient in intransitive sentences like those in (16b) and (17). Van Valin and LaPolla (1997) use "Undergoer" systematically instead of "Patient."

[6] Many of these verbs are what recent tradition has called "unaccusative." My inclination is to treat them as syntactically intransitive verbs like any other, just semantically special in their macrorole tier. But the amount of literature on their special syntactic properties (e.g. Levin and Rappaport Hovav 1995 and references therein) is beyond the scope of evaluation in this book.

Y', whose arguments are Actor and Patient, respectively.[7] It allows three configurations: both Actor and Patient, or either role alone. This function coexists with the standard expressions of thematic roles and constitutes the macrorole tier. (18) shows representative structures for some of the sentences above; the arguments of AFF are the macroroles.

(18) a. Bill threw the ball.

$$\begin{bmatrix} \text{BILL CAUSE [BALL MOVE]} \\ \text{BILL AFF}^{-} \text{ BALL} \end{bmatrix}$$

b. The car hit the tree.

$$\begin{bmatrix} \text{CAR MOVE TO CONTACT WITH TREE} \\ \text{CAR AFF}^{-} \text{ TREE} \end{bmatrix}$$

c. Bill loaded the truck with the books.

$$\begin{bmatrix} \text{BILL CAUSE [BOOKS MOVE ONTO TRUCK]} \\ \text{BILL AFF}^{-} \text{ TRUCK} \end{bmatrix}$$

d. Bill entered the room.

$$\begin{bmatrix} \text{BILL GO INTO ROOM} \\ \text{BILL AFF}^{-} \end{bmatrix} \quad \text{(no Patient)}$$

e. Bill died.

$$\begin{bmatrix} \text{BILL DIE} \\ \text{AFF}^{-} \text{ BILL} \end{bmatrix} \quad \text{(no Actor)}$$

f. The chocolate melted.

$$\begin{bmatrix} \text{CHOCOLATE BECOME LIQUID} \\ \text{CHOCOLATE AFF}^{-} \end{bmatrix}$$

(*or* AFF^{-} CHOCOLATE)

Note that in (18a,c), the direct object is a Patient on the macrorole tier. However, the direct object is not an argument of the main function CAUSE on the thematic tier; rather, it is an argument of the effect of causation. Thus what is caused in (18c) is that the books go on the truck, but Bill's action is conceptualized as being directed at the truck. This illustrates a virtue of the division into tiers: it allows conceptual structure simultaneously to express the overall effects of an action as well as the particular character at whom the action is directed.

[7] Other theorists have notated Actions in terms of a special function X DO... ('X does the action'); the intended distinction is the same. However, AFF is a bit more complex than DO, because it potentially has a second argument, filled by either a Patient or, as we will see shortly, a Beneficiary.

The reason for the minus sign superscripted to AFF in (18) is that there is a variant of the Patient role with somewhat parallel properties: Beneficiary. Whereas one does something *to* a Patient, one does something *for* a Beneficiary. The contrast is clear in a minimal pair like (19a,b). In (19c), the difference between Patient and Beneficiary is pragmatic, depending on whether meeting George is construed as a Good Thing or a Bad Thing.

(19) a. What Susan did to/*for Roberta was hurt her. (*Roberta* is Patient)
What Susan did to/*for Roberta was force her to leave.
b. What Susan did for/*to Roberta was help her. (*Roberta* is Beneficiary)
What Susan did for/*to Roberta was enable her to leave.
c. What Susan did for/to Roberta was introduce her to George.

One standard position for the Beneficiary role is indirect object position, as Recipient of an act of giving or an act of creation (the latter is the so-called *for*-dative).

(20) a. What Susan did for Roberta was give her a present. (= give a present to her)
b. What Susan did for Roberta was fix her a drink. (= fix a drink for her)

The difference between Patient and Beneficiary is notated by the choice of sign on AFF: if AFF$^-$, the second argument is Patient (negatively affected); if AFF$^+$, the second argument is Beneficiary (positively affected). Like Patient, Beneficiary can appear with an Actor (21a) or without (21b).

(21) a. Susan gave Roberta a present.

$$\begin{bmatrix} \text{SUSAN CAUSE [ROBERTA RECEIVE PRESENT]} \\ \text{SUSAN AFF}^{+}\text{ ROBERTA} \end{bmatrix}$$

b. Roberta received a present.

$$\begin{bmatrix} \text{ROBERTA RECEIVE PRESENT} \\ \text{AFF}^{+}\text{ ROBERTA} \end{bmatrix}$$

The macrorole tier plays an important role in linking conceptual structure to syntax. The basic idea stems from the insight that the relation between semantic roles and syntactic positions is not random. In particular, the Actor role, if present, has a strong claim on subject position (in active sentences). The general principle is an interface rule that can be stated informally as (22). (This is the version from Jackendoff 1990; similar principles appear as "linking hierarchies" in Anderson 1977, Bresnan and Kanerva 1989, Grimshaw 1990, Dowty 1991,[8] and Van Valin and LaPolla 1997, among others; see Culicover and Jackendoff 2005, chaps. 5 and 6, for refinement. For a proposal that there are multiple linking hierarchies from which individual languages can choose in various ways, see Aissen 1999.)

(22) *Linking of macroroles to syntax*

a. The first macrorole (Actor if there is one, otherwise Patient/ Beneficiary) is expressed in subject position.

b. The second macrorole, if there is one, is (canonically) expressed as the postverbal NP (indirect or direct object).

c. Any remaining NP arguments in the syntax (e.g. the direct objects of *enter* and *receive*) are linked to roles in the thematic tier.

There are also auxiliary interface principles that connect Patient and Beneficiary roles to syntactic positions. Two prominent cases are adversative and benefactive adjuncts, shown in (23).

[8] Dowty (1991) argues that the traditional semantic notion of agency actually has a number of subcomponents that can appear independently and that contribute independently to the likelihood that an NP will appear in subject position. The factors contributing to what he calls "Proto-Agency" are listed in (i)–(v) (Dowty 1991, 572).

(i) Volitional involvement in the event or state

(ii) Sentience (and/or perception)

(iii) Causing an event or change of state in another participant

(iv) Movement (relative to the position of another participant)

(v) Existence independent of the event named by the verb (possibly)

Of these, (v) probably has something to do with Topic, rather than agent, as Dowty notes. Criterion (iv) picks out the role here called "theme" (the object whose motion or location is being specified). Because theme precedes location in the linking hierarchy (Jackendoff 1990, chap. 11), themehood creates a pressure toward subjecthood if there is no agent. Criterion (iii) is the first argument of CAUSE (what I am calling here "agent"). The attribution of sentience (ii) appears when the character in question holds a situational or actional attitude in the sense of chapter 8 (however, I will call this case into question later in this chapter). Volitional involvement (i) is a subcase of holding an actional attitude. Dowty does not say how these factors are encoded in semantic structure; we have seen here that the relevant cases are all encoded in terms of structural positions as arguments of particular functions in conceptual structure. Thus his account falls out of the present one rather nicely.

(23) a. *Adversative adjunct*
My car broke down *on me.*
b. *Benefactive adjunct*
Amy fed the cats *for me.*

Although the arguments are clear enough, since 1990 I have been somewhat uneasy with the analysis, in that the macrorole tier contains only one possible function, AFF, in positive and negative variants. To pull its theoretical weight, the tier should allow more varied possibilities for content. The next section develops a direction that appears promising.

6.3 Experiencers and Stimuli

Informal discussions of lexical semantics always include the roles Experiencer and Stimulus in the context of verbs such as *fear*: the subject of *fear* is an Experiencer and the direct object is the Stimulus, the thing that causes the experience. In traditions in which semantic roles are simply named without analysis, this is fine. But in Conceptual Semantics, individuals get their semantic roles by virtue of occupying particular argument positions of semantic functions. For example, as illustrated in the previous sections, agent is the first argument of CAUSE in the thematic tier, effect is the second argument of CAUSE, and theme is the first argument of GO or BE. In the macrorole tier, Actor is the first argument of AFF, and Patient and Beneficiary are the second argument of AFF^{-} and AFF^{+}, respectively. To provide a similar account of Experiencer and Stimulus, I would like to experiment with introducing a new macrorole tier function *X EXP Y*, ‘X experiences Y’, in which the first argument is Experiencer and the second is Stimulus.

So let us finally turn to perception verbs. Consider the question of how to differentiate *look* and *see*. The subject of *look at* is an Actor, but that of *see* is not (24a); and *look at* can occur in the progressive, characteristic of actions, whereas *see* cannot (24b) (barring certain special pragmatic situations such as *I must be seeing things*).

(24) a. What I did was look at/*see the tree.
b. I am looking at/*seeing the tree.

Yet the thematic roles appear to be the same: I am making visual contact with the tree. The macrorole tier offers an option. Suppose that the thematic tier of both verbs has the function $X\ SENSE_{visual}\ Y$, ‘X senses Y in the visual modality’, which captures their commonality. Then *look at* could have AFF in the macrorole tier, and *see* could have EXP, as in (25).

(25) a. X looks at Y.

$$\begin{bmatrix} \text{X SENSE}_{\text{visual}}\ \text{Y} \\ \text{X AFF} \end{bmatrix}$$

b. X sees Y.

$$\begin{bmatrix} \text{X SENSE}_{\text{visual}}\ \text{Y} \\ \text{X EXP Y} \end{bmatrix}$$

At least four differences arise from this distinction.

- EXP makes the sentence stative, as seen in (24b): the sentence describes an experience rather than an activity. By contrast, AFF makes the sentence a standard activity.
- One can *look around* without looking at anything in particular; that is, *look* does not require a second argument. By contrast, one cannot *see* without seeing something; that is, *see* does require a second argument. (Even in the intransitive sentence *I can see*, the implication is that I can see *something*.)
- Even when one is looking at something, it is not a Patient: **What happened to the tree was Bill looked at it*. This difference is reflected in the macrorole tier, where AFF does not mark Y as a Patient but EXP does mark Y as a Stimulus.
- EXP allows the possibility of error or misdescription on the part of the Experiencer, for instance *In the perception experiment, Sam saw three dots, even though there were only two*. This possibility is unavailable with *look*: **Sam looked at three dots, even though there were only two*. This last difference marks *see* as a mental verb like *believe* and *intend* (chapter 8).

The posited distinction between *look* and *see* is also appropriate for *listen to* and *hear*, just by changing the modality of SENSE to *auditory*. More interesting are *taste*, *smell*, and *feel*, which have variants of both sorts. On the present analysis, active tasting, smelling (i.e. sniffing), and feeling (i.e. palpating) have AFF in the macrorole tier; and passive tasting, smelling, and feeling have EXP.

(26) a. Sam is carefully feeling the rug (for defects).

$$\begin{bmatrix} \text{SAM SENSE}_{\text{tactile}}\ \text{RUG} \\ \text{SAM AFF} \end{bmatrix}$$

b. Sam feels the rug (under his feet).

$$\begin{bmatrix} \text{SAM SENSE}_{\text{tactile}}\ \text{RUG} \\ \text{SAM EXP RUG} \end{bmatrix}$$

The verbs *sense* and *notice* ('come to sense') and the adjective *aware of* also express the function SENSE, but they leave the modality open. All have EXP in the macrorole tier.

6.4 AFF, EXP, and Theory of Mind

The distinction between AFF and EXP is important for another, nonlinguistic reason as well. Backing off from the formalism for a moment, recall the discussion of section 3.3. There I proposed that the character of awareness is determined in part by valuation features, which give percepts their "feel." Among these are the features [±external], which distinguishes percepts from images, and [±self-initiated], which distinguishes self-initiated from non-self-initiated experiences. In this light, the combination SENSE+EXP appears to encode the conceptualization of the "feel" of one's relation to a percept, which has the feature combination [+external, −self-initiated]. The subscript on SENSE picks out which "vertical" modality is responsible for the percept in question.

However, SENSE+EXP can be used not just to refer to one's own perception: it is also the means by which we attribute perception to others. In other words, this combination of functions is part of theory of mind. By contrast, SENSE+AFF yields the verb *look*, which denotes an observable action. It is possible to determine what someone is looking at without knowing whether he or she is seeing (i.e. experiencing) it. Thus the crucial function for describing an experience is EXP.

An organism that lacks this predicate may still feel committed to the reality of percepts (or have an experience of reality), because the valuation features are present in any event. But such an organism neither attributes nor denies such percepts to others, because it lacks the necessary concept to make such attributions.

We can now bring this observation to bear on discussions of theory of mind. The notion of theory of mind entered the literature with Premack and Woodruff's (1978) article, "Does the Chimpanzee Have a Theory of Mind?" Experiments have shown (Povinelli 2000; Tomasello, Call, and Hare 2003) that chimpanzees follow the gaze of others, but there is some dispute about whether they can connect someone's direction of gaze to his or her state of knowledge. The dispute is often phrased in terms of whether chimps understand that "seeing is knowing." But what we have just done suggests a more accurate way to phrase the question: it is whether chimps understand that "looking is seeing." Under the analysis just proposed, to look at something is to direct one's gaze to it—an action observable by others—which chimps *do* appreciate. To *see* something, on the other hand, is to experience it visually—an unobservable state. So the dispute about chimps' capabilities translates directly into the question of whether chimps can attribute visual experience to others—formally, whether they have the function EXP in their conceptual repertoire.

In other words, the function EXP is a predicate that constitutes a part of theory of mind. Again, lacking the function EXP in one's repertoire would not preclude one's *having* perceptual experiences. But it would preclude one's attributing experiences to others, and it would preclude *thinking* or *reasoning* about one's own experiences as well as those of others—which is precisely what theory of mind is supposed to be about.

For humans, of course, the natural assumption is that a person who is looking at something is visually experiencing it, and vice versa. However, we also recognize exceptions. One can look at something without actually seeing it; conversely, one can have visual hallucinations without looking at anything. Thus we can write a rule of defeasible inference along the lines of (27).

$$(27)\quad \begin{bmatrix} \text{X SENSE}_{\text{visual}}\ \text{Y} \\ \text{X AFF} \end{bmatrix} \Leftrightarrow_{\text{default}} \begin{bmatrix} \text{X SENSE}_{\text{visual}}\ \text{Y} \\ \text{X EXP Y} \end{bmatrix}$$

This is the rule "Looking is seeing." If chimps lack it, it may be because they lack the right-hand expression altogether.

According to this analysis, it does not make a lot of sense to think of theory of mind as "modular" in any sense remotely close to the way the term "module" is commonly used in the literature. Rather, theory of mind arises simply from having additional predicates such as EXP in the level of conceptual structure—new ways of construing experience. In particular, EXP is so tightly integrated into the inner workings of the formalism that it is pointless to think of it as "informationally encapsulated" after the fashion of Fodorian modules (Fodor 1983) or my own "structure-constrained" modules (Jackendoff 2002a, chap. 7). It's just an extra "gimmick" in the existing module of conceptual structure, which permits a whole new range of concepts and inferences to be constructed. (An analogy: imagine how the range of possible chemical compounds would be reduced if there were no such things as chlorine and boron.)

Subsequent chapters will introduce other functions that are conceptualizations of valuation features in conscious experience; some of these too contribute to theory of mind.

6.5 The Mapping of EXP to Syntax

Back to the trenches. Let's next examine another syntactic frame of *look*, *taste*, *smell*, and *feel*, shown in (28a); the auditory counterpart is the verb *sound* (rather than *listen* or *hear*), as in (28b).

(28) a. This looks/tastes/smells/feels wonderful to Sam.
b. This sounds wonderful to Sam.

These present two problems: what is the macrorole tier, and what is the thematic tier? The sentences are all stative, so the macrorole tier should contain EXP, as in (29). However, the relation of (29) to the syntax is curious, in that the order of Experiencer and Stimulus is the opposite of their order in (25) and (26). Let us put this problem off for a moment.

(29) SAM EXP THIS

Next, what is the thematic tier? The sentences attribute *wonderful* to *this*, at least in Sam's mind, so we need something along the lines of (30a) as part of the structure. However, (30a) combined with the macrorole tier (29) is not enough, as it does not specify the perceptual modality that distinguishes each of the verbs in (28) from the others. Given that these verbs are our perception verbs again, it would be nice if we could reuse the function SENSE in this frame. So let us tentatively adopt the structure (30b), in which what Sam senses is the attribution of wonderfulness to *this*, and in which the macrorole tier is (29), as desired.

(30) a. THIS BE WONDERFUL
b. This looks/sounds/etc. wonderful to Sam.

$$\begin{bmatrix} \text{SAM SENSE}_{\text{visual/auditory/etc.}}\ \text{[THIS BE WONDERFUL]} \\ \text{SAM EXP THIS} \end{bmatrix}$$

Understanding the structure in (30b) calls for some care. Notice first that these sentences share the subjective characteristic of other EXP sentences, in that Sam's judgment may be nonveridical: we can easily say *This looks wonderful to Sam, but it's really not.*

Second, notice that unlike what we found in (25)–(26), the second argument of EXP is not identical with the second argument of SENSE. We encountered a parallel situation with action verbs in examples such as *Bill loaded the truck with books* (18c): the caused action is that the books go on the truck, but the Patient (the character being acted upon) is just *the truck*. Similarly, in (30b), Sam senses the whole situation but experiences it in terms of a particular object. We will look at other cases like this below.

Next, we must be especially careful to distinguish (28a,b) from (31a,b), which mean something different. (In turn, (31a,b) are not entirely parallel, as shown by the continuations: *see* must be veridical in this context, but *hear* and *feel* need not be.)

(31) a. Sam sees that this is wonderful (—*but he's wrong).
b. Sam hears/feels that this is wonderful (—but he's wrong).

The tensed *that*-clauses suggest that (31a,b) are expressions of propositional attitude: Sam has an experiental relation to the truth of a proposition that describes a situation. By contrast, (28) expresses a direct experience of a situation, unmediated by a proposition. In addition, because *this* is buried in the subordinate clause in (31), it does not function as Stimulus, as it does in (28).

A closer paraphrase to (28) uses the verb *find*, which puts *Sam* in subject position but leaves open the modality of experience; this too can be nonveridical.

(32) Sam finds this wonderful (but it's not).

$$\begin{bmatrix}\text{SAM SENSE [THIS BE WONDERFUL]}\\ \text{SAM EXP THIS}\end{bmatrix}$$

But now we come face to face with the problem of mapping the macrorole tier into syntax. A very close paraphrase to (32) is (33), which has the Experiencer and Stimulus in opposite syntactic positions. (34a,b) are another well-known minimal pair, very close in meaning to (32)–(33).

(33) This seems wonderful to Sam.

(34) a. Sam regards this as wonderful.
b. This strikes Sam as wonderful.

The verbs *regard* and *strike* are representative of a sizable class of predicates whose syntactic properties have been discussed in the literature for years, dating back at least to Chomsky 1965. Chomsky observes that pairs such as (34a,b) are similar in meaning, noting in particular that both involve *Sam* in the role (now called) Experiencer and *this* in the role Stimulus, but in opposite grammatical positions. This is a serious problem because it contradicts one of the deepest assumptions of generative grammar, dating back to the earliest work: that underlying syntactic form reflects semantics, especially thematic roles, uniformly.[9] This assumption is confirmed by the observations at the outset of this chapter: there are no verbs that reverse the arguments of *enter*, *eat*, and *kill*. Pairs like those in (32)–(34) appear to be counterexamples to this fundamental position.

Linguists have taken two basic approaches to the problem. Both accept the standard assumption about the relation of underlying syntactic form to thematic roles. The first accepts that (34a,b) are synonymous and derives them both from a common underlying form. For example, Lakoff (1970) and Postal (1971) propose that the pattern in (34b) is derived transformationally from an underlying form with the phrasal order of (34a). Postal in particular seeks to explain certain odd syntactic characteristics of the (34b) pattern on the basis of its having undergone this derivation (see below). Belletti and Rizzi (1988) develop a similar account, using quite different theoretical machinery. However, so far as I know, these approaches never offer an explanation of why these particular lexical items, all with a particular kind of meaning, undergo this sort of derivation.

The second approach to these alternations—taken by, for example, Grimshaw (1990) and Pesetsky (1995), whose proposal will be discussed further in section 7.5—claims that (34a,b) are *different* in meaning; that is, either *Sam* or *this* has different thematic roles in the two sentences. From this difference in thematic roles comes a different linking to syntax. But because these proposals do not include an articulated theory of semantic structure, their claims cannot be adequately evaluated. It is all too easy to convince oneself of a delicate difference in meaning and, without formalizing it, talk oneself into it for the sake of its efficacy in deriving the syntax. The argument risks circularity: the theory predicts that there is a difference in meaning, so you go out and find one, no matter how shaky.

I am going to take a third tack here and claim that (34a,b) are indeed synonymous, apart from constructional aspects of meaning that apply to subject position, for example that subjects are more likely to be topics and (as we will see in chapter 8) can often be construed as volitional.

[9] This function of underlying syntactic structure was weakened in mainstream generative grammar in the late 1960s, but was reasserted as the Uniformity of Theta Assignment Hypothesis (Baker 1988) in the middle 1980s and as a property of Logical Form at about the same time. See Culicover and Jackendoff 2005, chaps. 2–3.

But I will not derive one from the other syntactically. I will instead attribute the difference in syntactic structure to a different linking between semantics and syntax. The idea is this: in the case of action sentences, there is a strong canonical mapping between macroroles and syntax, as laid out in (22). In particular, the Actor has a very strong claim on syntactic prominence and hence always grabs the subject position. On the other hand, the Experiencer-Stimulus dyad is less stable in terms of prominence, so there is no predetermined canonical mapping from the arguments of EXP to syntactic positions.[10] Consequently, each verb that has EXP in its meaning must individually specify which macrorole is mapped to subject.

If we notate the macrorole destined to be syntactically prominent by underlining, we can show the difference between *find* and *seem* as follows:

(35) a. Sam finds this wonderful.

$$\begin{bmatrix}\text{SAM SENSE [THIS BE WONDERFUL]}\\ \underline{\text{SAM}}\text{ EXP THIS}\end{bmatrix}$$

(also *Sam regards this as wonderful*)

b. This seems wonderful to Sam.

$$\begin{bmatrix}\text{SAM SENSE [THIS BE WONDERFUL]}\\ \text{SAM EXP }\underline{\text{THIS}}\end{bmatrix}$$

(also *This strikes Sam as wonderful*)

There are some grammatical reflections of this instability, especially in the paradigm with Stimulus subject. Postal (1971) notes that reflexives in object position with (at least some of) the Stimulus-subject verbs are odd (36a,b), whereas parallel Experiencer-subject verbs (36c) and Action verbs (36d) allow reflexives without difficulty.

(36) a. ?Sam strikes/impresses himself as pompous.
b. ?Sam smells funny to himself.
c. Sam regards himself as pompous. (Experiencer subject)
d. Sam smelled himself to see if he needed a shower. (Action: AFF rather than EXP)

[10] Such an instability might arise from a conflict between two factors involved in weighting for syntactic prominence (along the lines of Dowty's (1991) feature analysis of the macroroles; see note 8). One factor is which character is having an effect on which: the Actor is having an effect on the Patient, and the Stimulus is having an effect on the Experiencer. The other factor is (prototypical) animacy: the prototypical Actor is a volitional Actor, hence animate; the Experiencer is always animate. In the Actor-Patient dyad, these two factors strongly converge in favor of the Actor; in the Experiencer-Stimulus dyad, they are in conflict.

Stimulus-subject verbs are also ungrammatical in the passive (37a), whereas parallel Experiencer-subject verbs allow passive (37b,c) (see Culicover and Jackendoff 2005, chap. 6, for a possible account in the parallel architecture of chapter 2).

(37) a. *Harry is struck/impressed by Sam as pompous.
b. Harry is regarded by Sam as pompous.
c. Harry is often found pompous (by uninformed people).

The instability of the Stimulus-subject configuration also shows up in aphasia. It is well-known that agrammatic aphasics have difficulty interpreting passives like *The lion was chased by the tiger*, getting the characters in the right roles only at chance levels. Piñango (2000) shows that they have the same difficulty with Stimulus-subject sentences like *The boy pleases the girl*—though not with Experiencer-subject sentences such as *The girl likes the boy*.

Experiencer-subject verbs are not without their odd symptoms either: crosslinguistically, they often appear with dative subjects and even nominative objects (as in Icelandic (Yip, Maling, and Jackendoff 1987)). And, as we will see in chapter 7 (table 7.1), even in English there are considerably fewer Experiencer-subject verbs than Stimulus-subject verbs.

So both classes of Experiencer verbs are rather curious. I will not offer an account here of these oddities, but I will take them as an indication that the grammar is a bit uncomfortable about the mapping of EXP into syntax; things don't go exactly the way they ideally should.

Just to see the intricacy of lexical relations that can be generated by this little system, it's worth reviewing the readings of *look* and *see*.

(38) a. $\text{NP see NP} = \begin{bmatrix} \text{X SENSE}_{\text{visual}} \text{ Y} \\ \underline{\text{X}} \text{ EXP Y} \end{bmatrix}$

b. $\text{NP look at NP} = \begin{bmatrix} \text{X SENSE}_{\text{visual}} \text{ Y} \\ \text{X AFF} \end{bmatrix}$

c. $\text{NP look AP to NP} = \begin{bmatrix} \text{X SENSE}_{\text{visual}} \text{ [Y BE Z]} \\ \text{X EXP } \underline{\text{Y}} \end{bmatrix}$

We can also add two senses of *appear*, one active and one like the experiential sense of *look*.

(39) a. God appeared to Moses. (active)

$\begin{bmatrix} \text{MOSES SENSE}_{\text{visual}} \text{ GOD} \\ \text{GOD AFF (MOSES)} \end{bmatrix}$

b. God appeared immense to Moses.

$\begin{bmatrix} \text{MOSES SENSE}_{\text{visual}} \text{ [GOD BE IMMENSE]} \\ \text{MOSES EXP } \underline{\text{GOD}} \end{bmatrix}$

It is worth stressing how the solution proposed here for (32)–(34) goes against the grain of standard thinking in linguistic theory. The trend over the past 30 years has been to relieve individual verbs of the responsibility for determining the positions of their syntactic arguments, by proposing linking hierarchies along the lines of (22). Accepting a linking hierarchy as a linguistic universal entails that when we come up against apparent counterexamples such as *find*/*seem* or *regard*/*strike*, we have to find a semantic difference to account for the syntactic difference ("interface uniformity" in the sense of Culicover and Jackendoff 2005). The alternative proposed here is that this is *not* a universal of language: it pertains only to verbs expressing AFF—which were after all the verbs used to motivate the original argument.

The psychological predicates expressing EXP, on this view, are *genuine* counterexamples to the linking universals, and children *do* have to learn them one by one. Fortunately, they are learnable, given that children actually hear the relevant sentences that show them the right order of arguments. So in terms of learning, the position proposed here isn't too problematic.

Incidentally, another case that presents problems for a universal linking hierarchy is verbs of possession, where we find both orders of possessor and possessed.

(40) a. John has a book.
 b. The book belongs to John.

The choice between the two forms depends in part on factors like definiteness (**A book belongs to John*) but doesn't seem to have much to do with thematic roles. Again, these are not action sentences, although the possessor is some sort of Beneficiary.

6.6 Experiencer Verbs without Overt Experiencers

Next consider a minor syntactic variant on (28), in which no Experiencer is expressed.

(41) a. Pat looks/appears/sounds/feels/seems wonderful.
 b. The stew tastes/smells wonderful.

What semantic structure should be assigned to (41)? In particular, it would be desirable to keep SENSE as the function on the thematic tier, so that the differences of modality stay the same as in (28). However, we then require an Experiencer to fill the first argument of SENSE.

Who is the Experiencer, though? There are two possibilities. First, the speaker can be taken to be the implicit Experiencer; that is, the context can fill in *I*, to yield a structure parallel to (35b). But the Experiencer is not always understood to be the speaker. Consider *I heard from Joan that Pat looks wonderful.* Here the Experiencer might be Joan, or Joan might have heard it in turn from someone else. It seems to me that this sense conveys the effect of the stimulus on a nonspecific, generic observer. Such an individual can be expressed (in subject position) by the generic personal pronoun *man* in German and *on* in French, and sometimes by English *one*, *people*, and unstressed *you* (or *ya*; e.g. *Ya never see bubblegum commercials on TV any more*). I'll use the term *YA* in conceptual structure to stand for this generic individual. Following this line of reasoning, we get (42) as the thematic tier of (41) on the generic reading.

(42) Pat looks wonderful.
YA $\text{SENSE}_{\text{visual}}$ [PAT BE WONDERFUL]

Sentences with generic characters in them express generic situations. For instance, *People go to the movies on Saturday night* expresses not a particular event of someone going to the movies, but a generic event, in which *the movies* is not a particular movie and even *Saturday night* is not a particular Saturday night. On the other hand, a generic is subtly different from a universal quantification such as *Every person goes to the movies on Saturday night.* The generic is somehow less specific, less compulsive, more forgiving than universal quantification; it seems to convey a characteristic situation rather than an exhaustive person-by-person enumeration.

The same generic sense appears in (42). It's not as though, if you check every person, all of them find Pat to look wonderful, as in *Every person finds Pat to look wonderful.* Rather, the judgment is characteristic or typical of how people would find Pat to look. As a result, the wonderfulness comes to be a characteristic of Pat rather than a relation between Pat and some particular perceiver.

This use of YA as an implicit argument is not confined to psychological predicates. Consider the relation between *Sam is polite to his friends* and *Sam is polite.* One cannot be polite in the absence of social contact: politeness inherently requires other individuals to whom one is being polite. Therefore *Sam is polite* must have an implicit argument; but this argument is far more general than the one in *Sam is polite to someone*, though less specific than the one in *Sam is polite to everyone.* Rather, the appropriate implicit argument seems again to be the generic person YA: 'Sam is polite to people in general'. A predicate like *famous* also appears to incorporate YA. *Bill is famous* means something like 'People have heard of Bill'; you can't be famous if no one has heard of you.

Now let us think about the macrorole tier for (41). Since no Experiencer role is expressed, it might make sense to just omit this role from the macrorole tier. Such an analysis would parallel sentences with Patients but no Actors (e.g. (18e,f): *Bill died*; *The chocolate melted*). The result is a structure like (43).

(43) Pat looks wonderful.

$$\left[\begin{array}{l} \text{YA SENSE}_{\text{visual}}\text{ [PAT BE WONDERFUL]} \\ \text{EXP }\underline{\text{PAT}} \end{array}\right]$$

A different reading with the same syntax appears with the verb *feel*, as in *Pat feels wonderful*. On one reading, someone else is feeling Pat; this has a structure just like (43) but in the tactile modality. The other reading expresses Pat's own feelings, irrespective of modality (e.g. *I feel pretty!*). This comes out like (44), with no Stimulus picked out in the macrorole tier.

(44) Pat feels wonderful.

$$\left[\begin{array}{l} \text{PAT SENSE [PAT BE WONDERFUL]} \\ \underline{\text{PAT}}\text{ EXP} \end{array}\right]$$

An important aspect of (43) is that intuitively it presents itself as "observer-free" or "perspective-free"; it is not tied to any particular Experiencer. If I utter (43), I am not telling you who saw Pat and made a judgment. This is captured in the formalism by the use of YA as the first argument of SENSE and the absence of an Experiencer role with EXP.[11] The result is that (43) conveys a sense of being dispassionate or objective. What is curious about this is that the semantic structure of the sentence makes essential use of EXP, a quintessential theory-of-mind predicate; but at the same time the judgment is taken out of people's minds! This puzzle is a central concern of the next chapter, and it recurs in the discussion of value in chapter 9.

To sum up, this chapter has shown how different, subtly related senses of words can be described by combining a small number of functions in

[11] One can imagine notating this with other combinations, such as placing YA also in the macrorole tier or having no argument for the Experiencer in the thematic tier. I'm not sure what difference that would make, and for the moment I'm exploring the possibility that strikes me as most interesting.

semantic structure. For instance, we represent four different senses of *feel*—active palpation (26a), passive tactile sensation (26b), generic report of tactile properties ((43) in the tactile modality), and internal sensation of properties (44)—using combinations of the same primitive functions. At the same time, some of these senses are theory-of-mind verbs and others are not. We thereby see the way that theory of mind is woven into the human conceptual system, not as a module, but as a suite of functions within the conceptual repertoire.

Chapter 7

Objective and Subjective Psychological and Evaluative Predicates

7.1 The Problems

This chapter develops an analysis of psychological and evaluative predicates such as *interesting* and *bored*. Chapter 6 mentioned that, like some of the perception verbs, these constitute a domain that has historically presented difficulties for mapping semantic arguments to syntax. The problem is that (1a,b) seem close to synonymous, yet their subjects and objects are reversed.

(1) a. John fears sincerity. (Experiencer subject, Stimulus object)
b. Sincerity frightens John. (Stimulus subject, Experiencer object)

A more important issue from the standpoint of human conceptualization arises from the interpretation of these predicates. A good illustration comes from a rhetorical strategy all too familiar in academic circles: an authority figure utters something like (2),

(2) Problem P isn't interesting.

and everyone who works on problem P feels stupid and/or insulted, without quite knowing why. The rhetorical effect is considerably blunted, though, if the person in question instead utters something like this:

(3) a. Problem P doesn't interest me.
b. I'm not interested in problem P.

Intuitively, the difference is that (3) frames the interest as subjective, in the mind of the observer, whereas (2) frames being interesting as an objective or perspective-free property of the problem, about which it would be silly to disagree. Thus uttering (2) carries with it a subtext of presupposing power over the truth, rather than leaving room for differences of opinion.

The curious thing, of course, is that interest *cannot* inhere in a problem; it takes a *person* to be interested. So the problems are these: how does the conceptual system come to treat interest as an objective property of objects, and what is the difference between objective and subjective in this system of predicates? As observed in section 6.6, the same issue arises with perception verbs. For instance, in the sentence *You look marvelous, darling*, the verb is used as though the way you look is an objective fact, rather than depending on the eye of the beholder.

7.2 Classes of Affective/Evaluative Psychological Predicates

In order to tackle these two problems in a suitably general way, I want to look at the large and varied class of affective/evaluative terms of which *interesting* is a member. To tease apart what is systematic in the semantics from what is only partially systematic in the morphology, it helps to examine a substantial mass of linguistic detail.

Six grammatical forms are of interest here. In (4), these forms are arranged along a pragmatic cline, beginning with the form that most foregrounds the Experiencer and ending with the one that most foregrounds the Stimulus. Each stem can appear in only some of the forms, so I illustrate with both *bore* and *detest* in order to fill out the paradigm.

(4)	a.	I'm bored.		(Experiencer-Adjective)
	b.	I'm bored with this.		(Experiencer-Adjective-Stimulus)
	c.		I detest this.	(Experiencer-Verb-Stimulus)
	d.	This bores me.		(Stimulus-Verb-Experiencer)
	e.	This is boring to me.	This is detestable to me.	(Stimulus-Adjective-Experiencer)
	f.	This is boring.	This is detestable.	(Stimulus-Adjective)

Table 7.1 (pp. 220–223) enumerates about 70 such predicates, arranged by what forms they appear in.

A number of observations emerge from these data.

- The very same adjectives appear in frames (4e) and (4f). Hence the Experiencer argument in (4e) can be uniformly regarded as optional. This is reflected in table 7.1, where types (4e) and (4f) have been combined in column E/F.
- If a verb *V* occurs in the Stimulus-Verb-Experiencer frame (4d), it usually has a related Experiencer adjective *V-ed* (frames (4a,b)) and a re-

lated Stimulus adjective *V-ing* (frames (4e,f)). *Bore* is a typical example (*bored* and *boring*); such verbs appear in group Ia in table 7.1. However, some such verbs have adjectival derivatives with other affixes (e.g. *attract/attracted/attractive*, *disgrace/disgraced/disgraceful*, *endanger/endangered/dangerous*, *impress/impressed/impressive*, *nauseate/{nauseous/nauseated}/nauseating*, *offend/offended/offensive*, *scare/scared/scary*). Other verbs, found in group Ib, lack one of the related adjectives; *bug* (*This problem bugs me*) and *matter to* have no related adjectives.

- Many verbs that occur in the Experiencer-Verb-Stimulus frame (4c) have related Stimulus adjectives (frames (4e,f)), as seen in group II (e.g. *abhor/abhorrent*, *detest/detestable*, *enjoy/enjoyable*, *like/likable*, *loathe/loathsome*, *value/valuable*). But most such verbs lack related Experiencer adjectives (frames (4a,b)) (though the forms *fear/fearful/?fearsome* do exist).
- For the most part, the same verb does not occur in both the Experiencer-Verb-Stimulus frame (4c) and the Stimulus-Verb-Experiencer frame (4d). But Pesetsky (1995) points out some verbs that do (listed in group III): *delight*, *grieve*, *puzzle*, and *worry.*

(5) a. I delight in this. vs. This delights me.
b. I grieve over this. vs. This grieves me.
c. I've puzzled over this. vs. This puzzles me.
d. I worry about this. vs. This worries me.

- Like many adjectives in frames (4a,b,e,f), *bored*, *boring*, and *detestable* are morphological derivatives from verb stems. But in general this need not be the case, as seen in group IV. For instance, *happy*, *ecstatic*, and *nervous* appear in frames (4a,b) and are not related to verbs; and *funny*, *worthless*, and *important* appear in frames (4e,f). Still other patterns are possible, as seen in groups Ib and IV. For instance, the adjective *calm* appears in frame (4a) and reappears as the verb *calm* in frame (4d) and the adjective *calming* in (4e,f). *Elated* appears in frames (4a,b), but there is no verb *elate* (in my dialect) from which it can be "derived." *Apprehensive* appears in frames (4a,b), but the verb *apprehend* has a different meaning altogether. *Curious* and *sad* appear both as an Experiencer property (frames (4a,b)) and a Stimulus property (frames (4e,f)).
- The adjectives in frame (4b) differ in their choice of preposition. The choice is largely idiosyncratic: *bored* ***with*** *this*, *apprehensive* ***about*** *this*, *amazed* ***at*** *this*, and so on. There may be some degree of semantic motivation behind some of these choices, but I will not be concerned with this issue here.

Table 7.1
Psychological verbs and adjectives and their morphological variants

A	B	C
Exp-Adj (inherent)	Exp-Adj-Stim (directed)	Exp-Verb-Stim
I'm bored	I'm bored with this	I detest this
Group Ia: Stimulus-subject verbs with -ed *and* -ing *derivatives*		
	amazed at/about	
	amused at/about	
	annoyed at/about/with	
	attracted to	
bored	bored with	
depressed	depressed about	
	disgraced by	
	disgusted with/by	
distressed	distressed about	
	embarrassed about	
excited	excited about	
frightened	frightened about	
	horrified about	
	humiliated by/about	
	impressed with	
	insulted by	
	interested in/by	
	irritated with/by/about	
	moved by	
nauseated/nauseous	nauseated by	
	offended by	
	outraged by/with	
	pleased about/with	
scared	scared about/by	
	soothed by	
	surprised at/by	
terrified	terrified by	
	thrilled at/with/by	
upset	upset at/with/by	
Group Ib: Stimulus-subject verbs lacking either -ed *or* -ing *derivatives*		
angry	angry at/with/about	
calm	calm about	
elated	elated about	
	endangered	

D Stim-Verb-Exp This bores me	E/F Stim-Adj-(Exp) This is boring (to me)
amaze	amazing
amuse	amusing
annoy	annoying
attract	attractive
bore	boring
depress	depressing
disgrace	disgraceful
disgust	disgusting
distress	distressing
embarrass	embarrassing
excite	exciting
frighten	frightening
horrify	horrifying/horrible
humiliate	humiliating
impress	impressive
insult	insulting
interest	interesting
irritate	irritating
move	moving
nauseate	nauseating
offend	offensive
outrage	outrageous
please	pleasing
scare	scary
soothe	soothing
surprise	surprising
terrify	terrifying
thrill	thrilling
upset	upsetting
anger	
appeal to	appealing
bug	
calm	calming
(elate?)	
endanger	dangerous

Table 7.1
(continued)

A	B	C
Exp-Adj (inherent)	Exp-Adj-Stim (directed)	Exp-Verb-Stim
I'm bored	I'm bored with this	I detest this
enraged	enraged about	
(peevish?)	peeved with	
Group II: Experiencer-subject verbs		
		abhor
		detest
		enjoy
fearful	fearful about	fear
		hate
		like
		loathe
		value
Group III: Verbs with both Stimulus-subject and Experiencer-subject forms		
	delighted about/with	delight in
	grieved about/over	grieve over
itchy		itch (intrans.)
	puzzled about	puzzle over
		smell (trans.)
	worried about	worry about
Group IV: Adjectives not morphologically related to verbs		
afraid	afraid of/about	
apprehensive	apprehensive about	
curious	curious about	
ecstatic	ecstatic about	
happy	happy about/at	
	(incredulous?)	
joyful		
nice		
nervous	nervous about	
sad	sad about	

D Stim-Verb-Exp This bores me	E/F Stim-Adj-(Exp) This is boring (to me)
enrage	
matter to	
peeve	
repel	repellent/repulsive
shame (?)	shameful
stink (intrans.)	stinky
	abhorrent
	detestable
	enjoyable
	(fearsome?)
	hateful
	likable
	loathsome
	valuable
delight	delightful
grieve	
itch (trans.)	itchy
puzzle	puzzling
smell (intrans.)	smelly
worry	(worrying)
	curious (?)
	funny
	important
	incredible
	nice
	sad
	worthless

- In some cases, two separate Stimulus complements are possible in frame (4b) (e.g. *angry with Bill about the party*, *irritated at Mary about the mistake*). Pesetsky (1995, 60) distinguishes these two complements as "Target" and "Subject Matter" roles; for the moment, I will not distinguish them (though I acknowledge that something is missed thereby).

In short, the psychological verbs and adjectives display the usual mix of semiregularity and irregularity characteristic of derivational morphology.

7.3 Experiencer-Subject Adjectives and Verbs

Let us now begin to formalize the frames in (4). The goal is to provide a natural account of the semantic relations among them. When there are morphological relations among the frames, these provide important clues, because morphologically related items often share a semantic core. We begin with the Experiencer-subject adjectives in frames (4a,b).

There is an important division among adjectives in the Experiencer-subject frame (4a). Some of them, such as *bored*, can express pure or inherent "feelings," as shown in (6a). But others, such as *interested*, always have an implicit Stimulus argument: one cannot be interested without being interested in *something*. Thus (6b) is infelicitous.

(6) a. I'm not bored with anything in particular, I'm just (plain) bored.
(also *calm*, *depressed*, *distressed*, *elated*, *enraged*, *excited*, *happy*, *joyful*, *nervous*, *sad*, *scared*, *terrified*, *upset*)

b. *I'm not interested in anything in particular, I'm just (plain) interested.
(also *amazed*, *amused*, *annoyed*, *ashamed*, *disgraced*, *disgusted*, *horrified*, *insulted*, *offended*, *outraged*, *pleased*, *puzzled*, *surprised*, *thrilled*)

This distinction is reflected in the composition of columns A and B in table 7.1. Column A contains only those adjectives that can express inherent feelings, like *bored*. Column B contains those that can express feelings directed at something, such as *interested*. Since all the adjectives that express inherent feelings can also express directed feelings, column B also contains all the adjectives in column A (with a few exceptions that have a Stimulus-subject variant, such as *itchy* and *nice*).

The contrast between these two classes parallels that found between the verbs *swallow* and *eat*.

(7) a. I didn't swallow anything, I just swallowed.
b. *I didn't eat anything, I just ate.

Following the treatment of Jackendoff 2002a, chap. 5, the action of eating invariably involves two characters, only one of which is obligatorily expressed in syntax. Thus, if one is eating, one is inevitably eating *something*. By contrast, *swallowing* has only one obligatory semantic argument, the Actor; the material swallowed (the Patient) is an optional character in the action. The same analysis can be applied to *bored* versus *interested*, as diagrammed in (8).

```
(8) a. NP  swallow  (NP)      NP is bored (with NP)
       |              |       |                 |
       X  SWALLOW   (Y)       X  BORED        (Y)
    b. NP  eat  (NP)          NP is interested (in NP)
       |          |           |                     |
       X  EAT     Y           X  INTERESTED         Y
```

In other words, being bored can be either a "pure feeling" or one connected to a particular stimulus. This appears true of all the predicates in column A of table 7.1. By contrast, being interested is directed toward a stimulus (as are the other predicates in column B), even if the stimulus is not named but only inferred from context.[1]

This difference between "pure feelings" and "directed feelings" does not appear to have anything to do with language. Rather, it appears to arise from the character of human experience. Research on cultural universals of emotion (Ekman and Davidson 1994) seems to show that certain aspects of experience can be characterized as "moods" or "pure emotions," independent of surroundings; these include being happy, sad, calm, nervous, scared, and upset. Others are intrinsically "directed emotions," such as being attracted, disgusted, interested, humiliated, or ashamed; these require connection to a stimulus in the environment (or in one's mind). However, the "pure emotions" can also be directed at or connected to some particular stimulus. Thus the semantic classification illustrated in (6) is not accidental; the classes are psychologically natural.

[1] The implicit argument of *interested* differs from that of *eat* in that it is definite: *I'm eating* is shorthand for *I'm eating **something***, while *I'm interested* is shorthand for *I'm interested in **it/that***. Definite implicit arguments also occur for instance with *know* and *remember*: *I know/remember* means 'I know/remember it/that', not 'I know/remember something' (Grimshaw 1990; Culicover and Jackendoff 2005, 176).

The standard formalization of adjectival predication within Conceptual Semantics is shown in (9). An adjective in the syntactic frame *NP be Adj* expresses a property of the individual denoted by the subject. The subscripts *i* and *j* in (9) indicate that the semantic argument position X corresponds to the syntactic subject, and the semantic argument position Y corresponds to the adjective phrase.[2] For the time being, let us set aside the macrorole tier; we return to it in section 7.6.

(9) Phonology/syntax: $\mathrm{NP_i}$ be $\mathrm{AP_j}$
Semantics: $\mathrm{X_i}$ BE $[_{\mathrm{Property}}\ \mathrm{Y_j}]$
Example: Sam is old = SAM BE OLD

We want the adjectives under analysis here to conform to this template. Accordingly, we will encode them all as properties, as in (10)–(11).

(10) *Experiencer-Adjective-(Stimulus) (inherent or directed feeling)*
a. Sam is bored.
SAM BE $[_{\mathrm{Property}}$ BORED]
b. Sam is bored with school.
SAM BE $[_{\mathrm{Property}}$ BORED (SCHOOL)]

[2] This formalization differs in spirit from a more traditional logical treatment of adjectival predication, which in the present notation would come out as shown in (i).

(i) Phonology/syntax: $\mathrm{NP_i}$ is $\mathrm{AP_j}$
Semantics: $\mathrm{Y_j}\ (\mathrm{X_i})$
Example: Sam is old = OLD (SAM)

The treatment in (i) views simple adjectives as one-place predicates that happen to require a semantically empty verb *be* for syntactic well-formedness. By contrast, the treatment in (9) views a simple adjective as a semantic constant that denotes a position in "property space." On this analysis, the verb *be* is contentful: it establishes the connection between the subject and the property, just as it establishes the connection between the subject and a spatial location in *Sam is in Pittsburgh*. In Jackendoff 1983, chap. 10, I argue that the treatment in (9) more closely reflects the syntactic argument structure of adjective phrases, which can never have an internal subject. The argument there also leads to a nice treatment of a phrase like *get older*, which turns out to denote a change in property space "in the direction toward OLD"; and it permits a natural analysis of complex adjective phrases such as *three years older than Harry*, whose semantics comes out entirely parallel to that of spatial expressions such as *three miles down the road from Harry*.

Readers who nevertheless favor the treatment in (i) are encouraged to translate. In particular, they will have to add one more argument to the semantic analysis of each adjective, corresponding to the subject. For example, *interested in Z* will have two arguments instead of one: Z plus the subject.

(11) *Experiencer-Adjective-Stimulus (directed feeling only)*
a. Sam is interested.
SAM BE $[_{\text{Property}}$ INTERESTED (Z)]
b. Sam is interested in school.
SAM BE $[_{\text{Property}}$ INTERESTED (SCHOOL)]

The difference between the two cases lies in (10a) versus (11a): in the latter, even when an overt NP complement is absent, there is an "understood" or "implicit" argument Z toward which interest is directed. This follows the analysis in (8). Thus the general structure for adjectives like *boring* (those in column A of table 7.1) is (12a), and that for adjectives like *interesting* (those in column B but not column A) is (12b). The subscript on F distinguishes the two classes.

(12) a. *Inherent or directed feelings* (e.g. *bored*, *calm*, *depressed*, *happy*)
$[_{\text{Property}}$ F_i $\langle(Z)\rangle$]
(where $\langle(Z)\rangle$ denotes an optional argument)
b. *Directed feelings* (e.g. *amazed*, *amused*, *interested*, *pleased*)
$[_{\text{Property}}$ F_d (Z)]

So far the analysis includes frames (4a) (Experiencer-Adjective) and (4b) (Experiencer-Adjective-Stimulus). Next let us extend it to frame (4c) (Experiencer-Verb-Stimulus): verbs such as *like*, *loathe*, and *detest*. Intuitively, these denote a feeling on the part of the subject, directed toward the object. Because they are verbs instead of adjectives, they need no verb *be*. The simplest treatment of these verbs is to say that they combine a directed feeling with the predication function BE, as in (13).

(13) a. *Transitive verbs with Experiencer subjects* (e.g. *adore*, *fear*, *hate*, *like*, *loathe*)
Phonology/syntax: NP_i V NP_j
Semantics: X_i BE $[_{\text{Property}}$ F_d (Z_j)]
b. *Verbs with Experiencer subjects and oblique Stimuli* (e.g. *delight in*, *grieve over*)
Phonology/syntax: NP_i V $[_{PP}$ P $NP_j]$
Semantics: X_i BE $[_{\text{Property}}$ F_d (Z_j)]

Thus the meanings of these verbs conform to the standard predication template (9), but in a different way from the meanings of adjectives: in addition to specifying the property, they "incorporate" the predication function BE, so there is no need for the linking verb *be* in syntax. This parallels the incorporation of (for example) GO and INTO into the verb *enter*, as discussed in section 6.1.2. Like *enter*, these verbs leave open two characters in the situation, which come to be expressed as the subject and object (or oblique object) of the verb, respectively.

7.4 Stimulus-Subject Adjectives

Next let us consider frame (4e), the adjectives with Stimulus subjects and an overt Experiencer (e.g. *Golf is interesting to Bob*). For a very good first approximation, the two sentences in (14) describe the same situation. The only difference lies in how the relationship between *Bob* and *golf* is expressed.

(14) a. Bob is interested in golf. (Experiencer subject)
b. Golf is interesting to Bob. (Stimulus subject)

(14a) is of course the form treated in (11b). We would like to get (14b) from something very close to it. One approach involves a type of paraphrase common in formal logic: 'Golf is such that Bob is interested in it'. This paraphrase contains the same predicate as (14a), 'interested in', but it predicates over the *second* argument of the variable instead of the first. Such a paraphrase relation can be achieved formally with so-called lambda-abstraction, as in (15). The notation λ*z* can be read informally as 'such that', and the bound variable *z* that serves as argument of INTERESTED can be read as the resumptive pronoun 'it'.

(15) Golf is interesting to Bob.
GOLF BE [$_{\text{Property}}$ λz [BOB BE [$_{\text{Property}}$ INTERESTED (z)]]]
'Golf is such that Bob is interested in it.'

Under this analysis, the general form for adjectives in frame (4e) is (16).[3]

(16) *Stimulus-Adjective-Experiencer* (e.g. *amazing to X*)
[$_{\text{Property}}$ λz [X BE [$_{\text{Property}}$ F$_{\text{i/d}}$ (z)]]]
'such that X has such-and-such a feeling about it'

[3] It would be logically possible instead to relate the frames in (14) the opposite way, taking the Stimulus properties as basic and deriving the Experiencer properties by lambda-abstraction from them. I think this would be the wrong approach, for two reasons. First, it would give us no account of the inherent feelings such as *bored*, which involve no Stimulus. Second, what makes these adjectives "psychological predicates" is that they are essentially about the effect of things on observers. Thus it makes sense to take as basic the observer's reaction—the feeling—and build the Stimulus properties around this, rather than the other way around.

Now consider our original problem example, which falls into frame (4f): *That's not interesting*. It differs from (4e) only in that it lacks an overt Experiencer. The difference entirely parallels the treatment of *look/sound/seem/*etc. in section 6.6, right down to the use of the preposition *to*.

(17) a. Pat looks wonderful to Bob. (report of subjective judgment)
b. Pat looks wonderful. (perspective-free)
c. Golf is interesting to Bob. (report of subjective judgment)
d. Golf is interesting. (perspective-free)

Thus it seems appropriate to adopt the same situation in this case: to attribute the interest to a nonspecific, generic observer, notated as YA in conceptual structure.[4] On this analysis, *Golf is interesting* comes out as (18), and the general form for adjectives in frame (4f) is (19).

(18) Golf is interesting =
$\text{GOLF BE } [_{\text{Property}}\ \lambda z\ [\text{YA BE } [_{\text{Property}}\ \text{INTERESTED } (z)]]]$
'Golf is such that one is/people are interested in it.'

(19) *"Objective" stimulus-subject adjective* (same adjectives as in (16) but without *to X*)
$[_{\text{Property}}\ \lambda z\ [\text{YA BE } [\text{F}_{\text{i/d}}\ (z)]]]$
'such that ya/people have such-and-such a feeling about it'

Another possible analysis will be suggested in section 7.6.

7.5 Stimulus-Subject Verbs

Finally, consider frame (4d), the Stimulus-subject verbs such as *interest* in *This interests me*. There are a couple of possible analyses. The simplest would be to treat *This interests me* as essentially synonymous with *This is interesting to me*. It would then have the informal paraphrase 'This is such that I am interested in it' and the formal structure (20a). But (20a) is formally redundant, in that the outer BE and the lambda-abstraction

[4] Lasersohn 2005, which came to my attention after this book went to press, offers a formal semantic analysis of a subset of the Stimulus-subject predicates discussed in this chapter. Lasersohn notices that the covert Experiencer of the "objective" form is often established pragmatically; for instance, *Was that fun?* usually means 'Was that fun for you?'—a fact not noted in the text here. However, Lasersohn rejects the position adopted here that the Experiencer can be the generic 'people' or YA. I think that the fuller range of predicates discussed here and in the previous chapter weighs against his conclusion.

logically cancel out. A logically equivalent formulation is the far simpler (20b), whose informal paraphrase is ‘I am interested in this’.

(20) Phonology/syntax: NP$_1$ interests NP$_2$
Semantics: a. X$_1$ BE [λz [Y$_2$ BE
‘NP$_1$ is such that NP$_2$ is
[INTERESTED (z)]]]
interested in it.’
b. Y$_2$ BE [INTERESTED (X$_1$)]
‘NP$_2$ is interested in NP$_1$.’

The only difference between (20a) and (20b) is that in (20a), the lambda-abstraction makes the Stimulus more prominent. We might see this as the semantic correlate of placing it in subject position. The next section proposes another way to achieve the same effect, using the macrorole tier along lines suggested in chapter 6.

However, first we have to rule out some other possibilities. Pesetsky (1995) proposes a different meaning for Stimulus-subject verbs. On his view, these verbs are all causative: *This interests me* means roughly ‘This causes me to be interested in it’. This meaning would be formalized in the present framework as (21).

(21) Phonology/syntax: NP$_1$ interests NP$_2$
Semantics: X$_1$ CAUSE [Y$_2$ BE [INTERESTED (X$_1$)]]
‘NP$_1$ causes NP$_2$ to be interested in NP$_1$.’

Such an analysis is especially plausible for some verbs in this class, for instance *attract* (which Pesetsky does not cite). This verb’s spatial sense, as in *The magnet attracted the iron*, means roughly ‘X causes Y to move toward X’.[5] The psychological sense has a parallel feel: *This problem attracts me* suggests that the problem exerts a psychological force on me that moves me toward engaging with it. *Enrage* and *embitter* likewise seem plausible candidates, on the strength of paraphrases like ‘cause to become enraged’ and ‘cause to become bitter’.

[5] Well, not quite, as can be seen from an example like *The magnet attracted the iron filings, but they didn’t move*. A more correct analysis involves not CAUSE but a variation called “CS^u” in Jackendoff 1990, chap. 7, following the analysis of “force-dynamics” in Talmy 1988. The difference between CAUSE and CS^u can be illustrated by the difference between *Bill forced Sam to leave* and *Bill pressured Sam to leave*. Both involve the application of force on Sam by Bill. But only in the former does Bill achieve his goal of getting Sam to leave; this is true causation, or CAUSE. In the latter case, we don’t know if Sam actually left, which is the proper inference from CS^u.

On the other hand, other verbs in the class are far less comfortable in a causative paraphrase. Pesetsky himself points out some Stimulus-subject verbs such as *This appeals to me* and *This matters to me*, for which a causative paraphrase is nonsense: *'This causes me to be appealed to/mattered to by it'. Similarly, think again about *interest*. By comparison with 'cause to become enraged', the causative paraphrase 'cause to become interested in' feels rather lame.

To be sure, certain contexts can induce causative readings for many verbs in this class, as seen in (22a,b).

(22) a. Will tried/intended to please Harry. ($\approx$ 'Will tried/intended to make Harry be pleased with him.')
b. In order to puzzle the cops, ... ($\approx$ 'In order to make the cops puzzled, ... ')

But that doesn't necessarily mean that these verbs are *always* causative. To see why not, consider stative predicates like *be quiet*. We don't want to say that such predicates are invariably agentive: a machine or an evening can be quiet without any agent in sight. But contexts similar to (22) clearly imply volitional control over behavior.

(23) a. Will tried/intended to be quiet.
b. In order to be quiet, ...

As we will see in chapter 8, this sense of volitional control comes from "coercion" induced by *try*, *intend*, and *in order to*, rather than from the predicate *be quiet* itself. A predicate that cannot possibly be under volitional control is impossible in these contexts.

(24) a. *Will tried/intended to be born two years later.
b. *In order to be born two years later, ...

This parallel with nonpsychological predicates suggests that the sense of agentivity in (22) is likewise a consequence of coercion.

My sense therefore is that many Stimulus-subject verbs have a simple stative reading in which *This Xs me* (e.g. *This interests/pleases/worries me*) is effectively synonymous with *I am Xed Prep this* (*I am interested in/pleased with/worried about this*). In other words, there are both inherently causative Stimulus-subject verbs such as *attract* and *enrage*, and inherently noncausative ones such as *interest*, *please*, and *appeal to*. The latter, however, can be coerced into causative readings in certain contexts such as (22).

Pesetsky observes that in an overtly causative sentence, using the verb *cause* or *make*, the agent need not be identical with the Stimulus, as seen in (25a). On the other hand, the simple verbs *anger* and *worry* in (25b) do not allow the expression of a Stimulus distinct from the agent.

(25) a. The article in the paper [*agent*] made Sam angry at the government [*Stimulus*].
Dan's behavior [*agent*] made Barbara worry about his sanity [*Stimulus*].
b. The article in the paper angered Sam (*at the government).
Dan's behavior worried Barbara (*about his sanity).

Pesetsky takes this observation to reflect a deep-seated generalization that cries out for explanation in terms of Universal Grammar. He therefore embarks on a long train of reasoning that leads him to a radical reformulation of syntactic theory. In fact, though, the ungrammaticality of the final PPs in (25b) is far from fundamental. Pesetsky himself points out that many verbs in this class do allow distinct agent and Stimulus (26), often forming minimal pairs with nearly synonymous verbs that do not (27).

(26) a. Nancy riled Fred up about their taxes.
b. The news got Sam down about his income.
c. The concert turned me on to Beethoven.
d. The article in the paper pissed Sam off at the government.
e. Dan interested Barbara in chess.

(27) a. Nancy irritated Fred (*about their taxes).
b. The news depressed Sam (*about his income).
c. The concert excited me (*about Beethoven).
d. The article in the paper angered Sam (*at the government).
e. Dan intrigued Barbara (*with chess).

Pesetsky ends up capturing this difference with an ad hoc syntactic/morphological feature concealed deep in the machinery. It seems simpler (and more natural for the learner) to capture it with a superficial difference in syntactic subcategorization, which is the position I will take here. (See Culicover and Jackendoff 2005 for extended argument against strategies like Pesetsky's.)

Moreover, the verb types in (26)–(27) are not the only variants on causative Stimulus-subject verbs.

- With some verbs, the Stimulus is understood as identical with the agent; this is certainly the case with *This attracts/repels me*, for instance.
- With others, the Stimulus is only defeasibly the same as the agent, as in Pesetsky's example *This article about heart disease worries me—and not just about my* ***own*** *health.*

- With still others, the verb can express causation of an inherent feeling, as in *The ghost story frightened/depressed me*, where I was not frightened of or depressed with the ghost story; rather, I was just frightened or depressed. In such cases, the subject is no longer a Stimulus; it is only an agent.

Thus, when a genuinely causative reading is possible, the status of the Stimulus varies from verb to verb.

Pesetsky, in an effort to show that in general the Stimulus need not be identical with the agent, cites the following two examples (1995, 57–58):

(28) a. *John worried about Mary's poor health, but Mary's poor health did not worry John.
b. (*)Mary's poor health worried John, but John did not worry about Mary's poor health.

He claims that (28a) is a contradiction but (28b) is not, in that Mary's poor health may have caused John to worry about something else. I personally concur with Zubizarreta (1988; cited by Pesetsky (1995, 300n52)) in rejecting Pesetsky's judgment: I find (28a,b) equally bad. The equivalence is even clearer if the verb is *attract* or *annoy*.

(29) a. *John was attracted to/annoyed with the dog, but the dog did not attract/annoy John.
b. *The dog attracted/annoyed John, but John was not attracted to/annoyed with the dog.

However, if we change the verb to *frighten*, there is a difference.

(30) a. *John was frightened by the news, but the news did not frighten John. (contradictory)
b. The news frightened John, but John was not frightened of/about the news. (noncontradictory)

Note that if *of/about* is changed to *by* in (30b), the sentence *is* contradictory, for now the second clause is the passive of the first. This variation among the Stimulus-subject verbs confirms the conclusion that there is a lot of lexical variation in whether they require the Stimulus to be identical with the agent.

We end up with the following situation for the Stimulus-subject verbs. The simple noncausative readings are in general like the analysis of *interest* above; (31a) shows the general form. (31b–d) give three variants on the causative reading, differing in the status of the Stimulus argument.

(31) *Stimulus-Verb-Experiencer*

a. *Verbs with Stimulus subjects, noncausative* (e.g. *appeal to*, *interest*, *please*)
Syntax: NP_1 V NP_2 *or* NP_1 V [$_{PP}$ P NP_2]
Semantics: Y_2 BE [F (X_1)]

b. *Verbs with agent subjects, Stimulus as extra argument* (e.g. (26))
Syntax: NP_1 V NP_2 [$_{PP}$ P NP_3]
Semantics: Z_1 CAUSE [Y_2 BE [F(X_3)]]

c. *Verbs with agent subjects, necessarily identical with Stimulus* (e.g. *attract*, *repel*)
Syntax: NP_1 V NP_2
Semantics: Y_1 CAUSE [X_2 BE [F (Y_1)]]

d. *Verbs with agent subjects, defeasibly identical with Stimulus, but Stimulus can be either different or absent* (e.g. *frighten*, *depress*, *excite*) (defeasible argument is indicated in italics)
Syntax: NP_1 V NP_2
Semantics: Y_1 CAUSE [X_2 BE [F $\langle(\mathit{Y}_1)\rangle$]]

Some of the causative verbs give rise to "causative" adjectives that can be used actively. A prominent example is *annoying*, which is very comfortable in progressive aspect, and which denotes acting in a manner calculated to cause annoyance (32a). Such a context is next to impossible with *astonishing* (32b).

(32) a. Harry is being annoying.
HARRY BE
[$_{\text{Property}}$ λx [x CAUSE [YA BE [ANNOYED (HARRY)]]]]
b. *Harry is being astonishing.

This further confirms the lexical variation among these predicates.

That takes care of the six frames in (4), using the thematic tier alone. Now let us go back and look at them again with the help of the macrorole tier.

7.6 Adding the Macrorole Tier

As observed in section 7.1, the problem with the psychological verbs originally arose when linguists noted minimal pairs like those in (33). The verbs in (33a,b) are ones we'll discuss in this section; (33c) is a case we dealt with in section 6.5.

(33) *Experiencer subjects* *Stimulus subjects*
a. John fears rejection. Rejection frightens John.
b. John likes golf. Golf pleases John.
c. John regards Sue as smart. Sue strikes John as smart.

Following the treatment of the previous three sections, we can write structures like these for (33a):

(34) a. John fears rejection.
JOHN BE $[_{\text{Property}}$ AFRAID (REJECTION)]
b. Rejection frightens John.
i. REJECTION BE $[_{\text{Property}}$ λz [JOHN BE [AFRAID (z)]]]
or
ii. JOHN BE $[_{\text{Property}}$ AFRAID (REJECTION)]

We found the choice between (34bi) and (34bii) problematic. On one hand, the lambda-abstraction in (34bi) is formally overcomplicated and functions just to get REJECTION on the outside of the expression so it can be linked to subject position. On the other hand, (34bii) is identical to (34a), so it is not clear why *fear* and *frighten* are two different verbs.

The treatment of the macrorole tier in chapter 6 offers a resolution. All the verbs in (33) describe John's state of mind, so they all contain the macrorole function EXP. Section 6.6 proposed that EXP, unlike AFF, does not inherently determine which macrorole is linked to subject position. Rather, each EXP verb must individually mark its subject. This is exactly what we need for the contrasts in (33).

(35) a. John fears rejection.

$$\begin{bmatrix} \text{JOHN BE [AFRAID (REJECTION)]} \\ \underline{\text{JOHN}}\ \text{EXP REJECTION} \end{bmatrix}$$

b. Rejection frightens John.

$$\begin{bmatrix} \text{JOHN BE [AFRAID (REJECTION)]} \\ \text{JOHN EXP}\ \underline{\text{REJECTION}} \end{bmatrix}$$

In other words, we can abandon the overly complex (34bi) and still capture the difference between the verbs.

The macrorole tier also helps with the causative versions of the Stimulus-subject verbs. When there is an overt agent, the macrorole tier has to include AFF, since the agent (a kind of Actor) is acting on someone (who is therefore a Patient). For the clearest case, we can contrast the noncausative and causative versions of *interest*.

(36) a. Golf interests Bob. (noncausative = ‘Bob is interested in golf’)

$$\begin{bmatrix}\text{BOB BE [INTERESTED (GOLF)]}\\ \text{BOB EXP }\underline{\text{GOLF}}\end{bmatrix}$$

b. That article interested Bob in golf. (causative)

$$\begin{bmatrix}\text{ARTICLE CAUSE [BOB BE [INTERESTED (GOLF)]]}\\ \text{ARTICLE AFF BOB}\end{bmatrix}$$

Similarly, the causation of an inherent feeling can be treated as ordinary causation with AFF.

(37) The story depressed me.

$$\begin{bmatrix}\text{STORY CAUSE [I BE DEPRESSED]}\\ \text{STORY AFF ME}\end{bmatrix}$$

When the agent and the Stimulus are identical, as in *Bill amazed me*, there are two possibilities. One is (38a), with the expected CAUSE. The other is a structure where the thematic tier is the same as in the noncausative (38b), but the macrorole tier is different, as in (38c).

(38) Bill amazed me.

a. $$\begin{bmatrix}\text{BILL CAUSE [I BE [AMAZED (BILL)]]}\\ \text{BILL AFF ME}\end{bmatrix}$$ (agentive reading)

b. $$\begin{bmatrix}\text{I BE [AMAZED (BILL)]}\\ \text{I EXP }\underline{\text{BILL}}\end{bmatrix}$$ (pure experiencer reading)

c. $$\begin{bmatrix}\text{I BE [AMAZED (BILL)]}\\ \text{BILL AFF ME}\end{bmatrix}$$ (alternative for agentive reading)

This would make the two readings of *amaze* very much like the agentive and experiencer readings of *feel* (chapter 6, (26)) or like *look* and *see*. The reading in (38a,c) is an action, so it has the passive counterpart *I was amazed* ***by*** *Bill.* The reading in (38b) instead has a counterpart with the Experiencer-subject adjective, *I was amazed* ***at*** *Bill.* I am not sure that this is the correct analysis and that CAUSE can be dispensed with here; but the possibility is intriguing.

So far we have assigned a macrorole tier to the psychological verbs. Now let us consider the adjectives. I am not sure whether they should have a macrorole tier at all. However, if they do, since they all describe mental states, the function should be EXP (with the exception of the "causative" adjectives like *annoying*, which we return to below). The Experiencer-subject adjectives are no problem.

(39) a. Sam is bored. (inherent feeling)

$$\begin{bmatrix}\text{SAM BE BORED}\\ \text{SAM EXP}\end{bmatrix}$$

b. Sam is amazed at Frank. (directed feeling)

$$\left[\begin{array}{l}\text{SAM BE [AMAZED (FRANK)]}\\ \underline{\text{SAM}}\text{ EXP FRANK}\end{array}\right]$$

For the Stimulus-subject adjectives, there are two possibilities. On one hand, we could just add an EXP tier to the configurations worked out in section 7.4 (40). On the other hand, we could eliminate the formally clumsy lambda-abstraction and do the work with the macrorole tier, as we did with the Stimulus-subject verbs (41).

(40) a. Frank is amazing to Sam.

$$\left[\begin{array}{l}\text{FRANK BE [}\lambda\text{z [SAM BE [AMAZED (z)]]]}\\ \text{SAM EXP }\underline{\text{FRANK}}\end{array}\right]$$

b. Frank is amazing.

$$\left[\begin{array}{l}\text{FRANK BE [}\lambda\text{z [YA BE [AMAZED (z)]]]}\\ \text{EXP FRANK}\end{array}\right]$$

(41) a. Frank is amazing to Sam.

$$\left[\begin{array}{l}\text{SAM BE [AMAZED (FRANK)]}\\ \text{SAM EXP }\underline{\text{FRANK}}\end{array}\right]$$

b. Frank is amazing.

$$\left[\begin{array}{l}\text{YA BE [AMAZED (FRANK)]}\\ \text{EXP FRANK}\end{array}\right]$$

The one exception to the use of EXP is with causative adjectives like *annoying*, which come out perhaps like (42).

(42) Frank is being annoying.

$$\left[\begin{array}{l}\text{FRANK BE [}\lambda\text{x [x CAUSE [YA BE [ANNOYED (x)]]]]}\\ \text{FRANK AFF}\end{array}\right]$$

The only possible surprises here are (39a) and (40b)/(41b). (39a) describes an experience of Sam's that has no connection to the world, so there is no Stimulus on the macrorole tier. (40b)/(41b), following the analysis of *Pat looks wonderful*, has no Experiencer on the macrorole tier. As a consequence, the sentence expresses a perspective-free, "objective" judgment of Frank's properties. And this—finally!!—is the result we have been seeking: the reason why *Problem P isn't interesting* presents itself as an objective judgment.

7.7 Valence in the Macrorole Tier: More Theory of Mind

Recall from section 6.2 that the macrorole function AFF comes with a valence. AFF^{-} renders its second argument a Patient, a character that the event happens *to*; AFF$^+$ renders its second argument a Beneficiary, a character on whose behalf the event occurs. The difference between the two is often lexically marked; for instance, *hurting* is AFF$^-$ and *helping* is AFF$^+$. In the context of the psychological predicates, the newly introduced macrorole function EXP has similar valence properties.

To see this, we begin with the commonplace observation that emotions come with positive or negative valence: *happy* and *calm* versus *sad*, *angry*, and *scared*. Thus part of the conceptual structure of inherent feelings ought to be a valence feature. Directed feelings have a valence too (with a few possible exceptions such as *impressed* and *puzzled*, for which I have difficulty making a judgment). Let's notate this with a plus or minus sign on the function.

(43) a. delighted about Z
[DELIGHT$^+$ (Z)]
b. disgusted with Z
[DISGUST$^-$ (Z)]

The valence of a directed feeling turns out to be reflected in the macrorole tier. This is clearest when the macrorole function is AFF, where we already know what valence means. Consider cases like causative *amuse* and *annoy*, which have analyses like (37) or (38a,c). When the feeling is of positive valence, as with *amuse*, the sentence is taken to benefit the Experiencer; when the feeling is negative, as with *annoy*, the sentence is taken to affect the Experiencer negatively. Thus we might say that AFF is "tuned" to the valence of the feeling.

(44) a. What Sue did for/*to Tim was amuse him. (AFF$^+$)
b. What Sue did to/*for Tim was annoy him. (AFF$^-$)

It makes sense to extend this tuning of valence to EXP.

(45) a. Sue is delighted with Tim.

$$\begin{bmatrix}\text{SUE BE [DELIGHTED}^+\text{ (TIM)]}\\ \underline{\text{SUE}}\text{ EXP}^+\text{ TIM}\end{bmatrix}$$

b. Sue is disgusted with Tim.

$$\begin{bmatrix}\text{SUE BE [DISGUSTED}^-\text{ (TIM)]}\\ \underline{\text{SUE}}\text{ EXP}^-\text{ TIM}\end{bmatrix}$$

We might think of EXP with a valence as 'Experiencer has a good/bad experience of Stimulus'. (Other cases of EXP, for instance with the verb *see*, have a neutral valence.)

The principle for tuning the macrorole function to the valence of a feeling can be stated as follows:

(46) *Tuning of valence*

$$\begin{bmatrix} [\ldots \mathrm{F}^{\alpha} \ldots] \\ (\mathrm{X})\ \mathrm{AFF/EXP}^{\alpha}\ (\mathrm{Y}) \end{bmatrix}$$

(where α ranges over + and −)

This is a well-formedness condition on conceptual structures that guarantees that the two valences match.[6]

Section 6.4 related EXP to the *valuation features* of consciousness proposed in chapter 3, the "feels" that are attached to entities in experience. In the context of perception verbs such as *see*, EXP serves as a conceptualization of the valuation [+external, −self-initiated], the "feel" of a percept. When signs of valence are attached to EXP in the context of evaluative predicates, another valuation feature comes into play. The Stimulus in EXP^{+} corresponds to a [+affective: valence+]-perceived entity, one that carries a positive emotional coloring. The Stimulus in EXP^{-} corresponds to a [+affective: valence−] entity, one that carries a negative emotional coloring. (The perception verbs, which lack valence on EXP, are [−affective] or neutral.)

As usual, valuation features appear only in one's own experience, that is, when the Experiencer is ME: one feels the Stimulus viscerally as pleasant or unpleasant. By contrast, one cannot experience someone else's valuation features. So the valence of EXP serves as conceptual proxy for other people's experience. It is therefore another aspect of theory of mind, a conceptualization that enables us to reason about others' experiences as though they are parallel to our own.

7.8 Why Subjective and Objective Systems?

Throughout this chapter, we have been distinguishing "subjective" and "objective" versions of evaluative and psychological predicates. Chapter 9 will extend this subjective/objective duality to values, including moral values. This is a key to the fact, remarked in chapter 5, that moral systems are conceived of as objective, universal, and timeless, and why the term "moral relativism" is taken by many to be self-contradictory or tantamount to "amoral."

Why should we have these distinct systems in cognition? And what does one have to do with the other? As observed at the beginning of the

[6] This is overly simple, in that *That's **not** interesting* should have *negative* valence, even though the directed feeling itself is positive. Let's put off this important technical detail for another occasion.

chapter, there is an important sense in which "subjective" evaluation is more true to life: being of interest or being boring is fundamentally a relation between an object and a perceiver. Yet experientially, "objective" evaluation is every bit as valid: we experience certain objects simply as repulsive and certain people simply as attractive, and this is not taken to be a fact about our perception of them.

In the case of an "objective" evaluation, one's own contribution to the evaluative judgment is completely transparent, just like one's own contribution to the judgment of an object's size or color. It is computationally simpler, since the property is invariant across perceivers. However, we also need the "subjective" system in order to conceptualize individual differences in evaluation: this book is exciting to you and boring to me, this action is detestable to you and appealing to me. However, it takes theory of mind (including of my own mind) to recognize these differences—always a cognitive stretch.

In practical reasoning, we jump readily between the two systems. Something like (47) seems to be the appropriate rule of inference.

(47) *Objectification and Subjectification*

$\text{Y BE } [_{\text{Property}}\ \lambda z\ [\text{X BE } [\text{F } (z)]]] \Leftrightarrow_{\text{default}}$

(e.g. *Y is interesting to X*)

$\text{Y BE } [_{\text{Property}}\ \lambda z\ [\text{YA BE } [\text{F } (z)]]]$

(e.g. *Y is interesting*)

(where Y is the entity being evaluated, X is the experiencer, and YA is the generic perceiver)

First let us read the rule from left to right. Suppose we take ME as the experiencer X. Then the rule says that if I like Y, or Y is interesting to me or valuable to me, then, other things being equal, Y is objectively interesting or valuable. In other words, my own judgments by default warrant a judgment of objective value. Alternatively, suppose I don't know anything about Y but I find out that you like it, or that it's interesting to you or valuable to you. Then, using YOU as the Experiencer in (47), I can conclude by default that Y is objectively good or valuable. In other words, from left to right, (47) represents the objectification of evaluation.

Why should it be important to arrive at an objective value? The reason is that then the rule can be read from right to left to predict someone else's reactions to Y. In general, I can't reliably predict your evaluation of an object without evidence about your reactions. In the absence of such evidence, I fall back on (47) to predict your reaction. If something is objectively interesting or valuable, then it's reasonable to believe it will be interesting or valuable to you and me and anyone else.

In practice, then, we slip between the two systems as convenient. We strongly prefer the more predictive objective system when possible, but we can easily drop into the subjective system when we have evidence of difference. This is hardly logical reasoning—but it's what we do.

However, there is a value judgment that arises from a conflict between the objective and subjective systems.

(48) Someone whose judgment conflicts with objective reality is of negative esteem.

We will not have the formal tools to make this principle more precise until chapter 9, but the idea should be clear for now: it's bad to be wrong.

(48) raises a difficult practical problem: when I disagree with you, the question arises as to who has control of objective value and truth. One possibility is that I trust your judgment, say because you're an authority figure. Then (48) leads to the conclusion that *I'm* bad, and my self-esteem goes down. Thus (48) gives us the final piece in the puzzle raised at the beginning of this chapter, of how *Problem P isn't interesting* can trigger self-doubt.[7]

The more standard situation, though, comes about when I understand *myself* to be in possession of objective truth and evaluation, and I thereby think less of *you*. This is typically the case when two cultures encounter one another and each characterizes the other as uncultured, savage, and lacking in values. As in chapter 5, there is no need to recount the unpleasant consequences—between generations, between religions, between religion and science, between the sciences and the humanities, even between subcultures of a discipline. Dialogue can only take place if both protagonists are capable of switching into the subjective system for their own judgments as well as for the other's. As suggested above, this switch is always a cognitive stretch. In addition, because it drops the presumption of one's own correctness, it carries the risk of losing self-esteem. Intellectual humility consists in part of the willingness to take such a risk.

[7] I find it an interesting and important question how I "repair" my knowledge base, replacing "Y is good/true" with "Y is bad/false, even though I used to believe the opposite." But this goes beyond the scope of the present enterprise.

Chapter 8

Intending and Volitional Action

8.1 Introduction

Much of the philosophical and psychological literature on the "folk theory of mind" speaks of it in terms of "propositional attitudes," which are characterized simply as "beliefs, desires, and the like."[1] However, in order to understand how people reason about the minds of others, we need a more highly differentiated account of the attitudes. The present chapter investigates one particular case, *intending*, with attention to related and contrasting cases.

I focus on intending rather than believing here because, as I will show, it has a somewhat more complex structure, which reveals more of the texture of the folk theory of mind. In particular, the notion of volitional action—of performing an action intentionally—is crucial for understanding others' minds, and it also has well-known grammatical repercussions (for some of these, see Culicover and Jackendoff 2005, chap. 12). Moreover, an analysis of intending and volitional action is fundamental for the treatment of all manner of social interaction. For instance, speech acts typically involve the speaker's intending the hearer to come to know something or intending to get the hearer to produce some response (see sections 8.6.3 and 8.7). Transactions involve each character's doing something for the other with the intention of getting something in return

[1] An earlier version of this chapter appeared in a festschrift for Carlos Otero (Jackendoff 1995); a few passages have also been borrowed from Jackendoff 1985. Carlos has been a dear friend since 1970, when his enthusiasm for my work and his gentle support on both substantive and political matters in the "linguistics wars" meant a great deal to me. I am further grateful to Carlos for his avid commentary on my preliminary attempts to lay out the domain of social cognition.

(section 10.5). And a basic aspect of cooperation is the notion of joint intention and joint action (sections 5.8, 8.8).

As in chapters 6 and 7, I am addressing the issues in terms of conceptual structure. I am concerned with what the word *intend* means, that is, how people conceptualize situations in which someone can be said to intend something. In doing so, I take myself to be studying a human concept, not an aspect of ultimate reality: I am not concerned with what is *really* going on in people's brains when we attribute intentions to them. This may have been clearer in the case of *look* and *see* in chapter 6. No one would want to claim that the conceptual structures assigned to these verbs bear any resemblance to what is really going on the visual system, but the analysis does capture important features of how we *conceptualize* looking and seeing. To put this differently, unlike Fodor (1987), I do not assume that the folk theory of mind need bear any resemblance to a scientific theory of mind.

However, I also disagree with Churchland (1981) and Stich (1983), who regard the folk theory of mind as "simply false" and therefore without scientific interest. A scientific theory of mind must describe the range of human concepts. Among these are concepts about our own minds and the minds of others (e.g. Dennett's (1987) intentional stance), along with folk concepts of space and force—regardless of how "correct" or "incorrect" such concepts are scientifically.

Much of my analysis here is based on discussions in Searle 1983 and Bratman 1987; Miller and Johnson-Laird 1976 has also been useful.

8.2 Animate Actions as a Special Class of Situations

To be able to characterize intending and volitional action, first we have to understand the linguistic data, and that in turn requires some groundwork. Let's start with the issue of reference: how language refers to the world. Returning to a point from section 6.1: in Conceptual Semantics, the reference of linguistic expressions is not tied directly to entities in the world, as in the semantic theories of Frege, Tarski, and modern formal semantics, as well as some cognitivists such as Fodor. Rather, expressions refer to *entities as conceptualized by the language user*. The "real world" of objects "out there" is given to the language user by the interaction of the perceptual system and the conceptual system, as discussed in chapters 2 and 3 (see also Jackendoff 1983, chap. 2; 2002a, chaps. 9 and 10). This situates the theory of meaning and reference squarely within psychology.

One type of expression that is very clearly referential is so-called *deictic anaphora*—pronouns whose reference is fixed not by some previous expression but by something in the visual environment, often aided by pointing. (1) illustrates with a point to an object.

(1) Would you pick that up right now [*pointing at a pencil on the floor*]?

But it is possible to point to entities other than objects. The construction *that... happen* can be used to point to any sort of event (2a), and the constructions *do that* and *do this* can be used to point to actions (2b,c).

(2) a. *That* [*pointing at someone lying drunk on the floor*] had better never happen in *my* house!
b. You'd better not do *that* around here [*gesturing at the addressee spitting*]!
c. Can you do *this* [*demonstrating a skateboarding trick*]?

A basic conclusion arising from these cases is that human perception and conceptualization parse the world not only into objects, but also into events and actions in which these objects take part. (This parsing has been studied experimentally by Cutting (1981) and Zacks and Tversky (2001), for example.) Moreover, situations and actions can be referred to not only with pronouns (as in (2)), but also with certain kinds of noun phrases (3) and with sentences (4).

(3) a. a performance of *Harold in Italy* by Zukerman on Thursday
b. the US invasion of Iraq
c. Bill's speech at the conference

(4) a. Zukerman played *Harold in Italy* on Thursday.
b. The US invaded Iraq.
c. Bill spoke at the conference.

Mainstream semantic theory, growing out of the Fregean tradition, usually takes sentences to refer to truth-values rather than to events and actions. However, consider the use of the pronoun *it* in examples like (5b), exactly parallel to (5a).

(5) a. There was a performance of *Harold in Italy* by Zukerman on Thursday. I heard *it*. *It* was fabulous.
b. Zukerman performed *Harold in Italy* on Thursday. I heard *it*. *It* was fabulous.

In (5b), I certainly did not hear the truth-value of the proposition that Zukerman performed, and I did not think this truth-value was fabulous! Rather, just as in (5a), I heard the *event*, and the *event* was fabulous. In

(5a), *it* corefers with *a performance* . . . , and in (5b), *it* corefers with the entire first sentence. Thus the sentence too must refer to the event. (Such evidence is rarely considered in the philosophical literature on reference, although it was a mainstay of the briefly fashionable theory of Situation Semantics (Barwise and Perry 1983).)

Events and actions fit into a larger hierarchy of *situations*, which also include states such as that expressed by *Bill is tall*. The standard linguistic criterion for distinguishing events from states is that events are things that *happen*, and states are not.

(6) a. *Events*
What happened was (that) Zukerman performed *Harold in Italy* on Thursday.
What happened was (that) the US invaded Iraq.
What happened was (that) Bill received a letter.
What happened was (that) Fred was hit by a falling brick.
b. *States*
*What happened was (that) Bill was tall.
*What happened was (that) I had a bicycle.
*What happened was (that) Sue liked ice cream.

Actions in turn are a subclass of events, in which an Actor (expressed by the subject of the sentence) can be said to *do* something. We saw this test in section 6.2 and used it to differentiate *look* (an action) from *see* (a state).

(7) a. *Actions*
What Zukerman did was perform *Harold in Italy* on Thursday.
What the US did was invade Iraq.
b. *Nonaction events*
*What Bill did was receive a letter.
*What Fred did was be hit by a falling brick.

An Actor need not be acting intentionally (8a) or even be *capable* of acting intentionally (8b).

(8) a. What Bill accidentally did was roll down the hill.
b. What the rock did was roll down the hill.

However, before the issue of intent or lack thereof can be raised, a sentence must express an action, with an animate Actor. For instance, neither *intentionally* nor *unintentionally* may appear with the states or nonaction events in (6b) and (7b); nor may they appear with actions with an inanimate Actor as in (8b).

(9) *Bill (un)intentionally received a letter.
*Bill (un)intentionally was tall.
*The rock (un)intentionally rolled down the hill.

I will use the term *animate actions* to refer to actions that are capable of being intentional.

We end up with the ontology in (10).

(10) Situations $\supset$ Events $\supset$ Actions $\supset$ Animate actions

8.3 Situational and Actional Attitudes

We next undertake a general exploration of the syntax and semantics of propositional attitude verbs. A primary distinction among propositional attitudes emerges clearly in the difference between *believing* and *intending*. A belief is an attitude one can adopt toward any situation (state or event), concrete or abstract, at any time, with any combination of characters in it. For example, the subordinate *that*-clauses in (11) express situations. (A similar range of subordinate complement clauses is possible with *claim*, *imagine*, *doubt*, *assume*, and *presume*, among many others.)

(11) John believed...
that he was shorter than Bob.
that Bob was born 10 years earlier than he really was.
that Susan was descended from royalty.
that the sky is green.
that Sue would bring a cake to the party.

By contrast, one can hold an intention only with respect to an action in which one is oneself the Actor—that is, a *self-initiated action*. Thus the standard syntactic structure that goes with *intend* is an infinitival verb phrase (VP) whose subject is understood to be the subject of *intend*. Since an intender is necessarily animate, the acceptable VPs after *intend* are all animate actions. (A similar range of possibilities occurs with *be willing*, *plan*, and *offer*, among others.)

(12) John intended...
to look at Sue.
to scratch his nose. (actions)
to prove Fermat's theorem.
*to be shorter than Bob.
*to have been born 10 years earlier. (nonactions)
*to be descended from royalty.

I will call *believe*, *imagine*, and so forth verbs of *situational attitude* and *intend*, *be willing*, *plan*, and so forth verbs of *actional attitude*.

In addition to being distinguished by the semantic difference illustrated in (11)–(12), actional attitudes are distinguished from situational attitudes by their time-dependence. For example, a belief and a claim can be directed toward a situation at any time, past, present, or future; but an intention cannot be directed toward an action in the past, as seen in (13).

(13) a. John believes himself to have talked to Sue yesterday.
b. John claims to have talked to Sue yesterday.
c. *John intends to talk/to have talked to Sue yesterday.

Future-directedness (or better, *nonpast-directedness*) occurs in all other actional attitudes as well.[2]

(14) $*\left\{\begin{array}{l}\text{John planned today}\\ \text{John decided today}\\ \text{Bill persuaded John today}\\ \text{John is willing today}\\ \text{Today it occurred to John}\\ \text{John is obliged}\\ \text{John swore today}\end{array}\right\}$ to talk/to have talked to Sue yesterday.

[2] However, future-directedness is not confined to actional attitudes. It also occurs with infinitival complements of the situational attitude verbs *wish*, *desire*, and *expect*.

(i) $*\left\{\begin{array}{l}\text{John wishes}\\ \text{John desires}\\ \text{John expects}\end{array}\right\}$ to talk/to have talked to Sue yesterday.

The *that*-complements of these verbs are all different, though. The conditional *that*-complement of *wish* can be past-directed, as seen in (iia). On the other hand, the subjunctive *that*-complement of *desire* cannot be past-directed (iib). *Expect* preserves its future-directedness with an indicative *that*-complement: (iic), if acceptable, conceals a future-directed coerced interpretation something like (iid) or (iie).

(ii) a. John wishes that he had talked to Sue yesterday.
b. *John desires that he talk/have talked to Sue yesterday.
c. ??John expects that Bill talked to Sue yesterday.
d. John expects *to find out* that Bill talked to Sue yesterday.
e. John expects *it to turn out* that Bill talked to Sue yesterday.

There is another difference between the two kinds of complements of *wish*. The infinitival complement carries a strong sense that it is contrary to fact. But it is not *necessarily* contrary to fact, as shown in (iiia). By contrast, the conditional *that*-complement *is* necessarily contrary to fact, as shown in (iiib).

(iii) a. I wish to be exactly as I am.
b. *I wish that I were exactly as I am.

Despite these differences, there are important parallels between situational and actional attitudes. For instance, one of the long-term philosophical and logical problems with propositional attitudes (here, situational attitudes) has been their so-called referential opacity.[3] There are three well-known symptoms of referential opacity. First, verbs of propositional attitude can be followed sensically by clauses that are contradictory in isolation, such as *Susan is taller than she is*; this is seen in (15a) (this symptom was pointed out by Bertrand Russell). Second, characters introduced within a propositional attitude, such as *a goat* in (15b), resist wide scope existential generalization; that is, the inference in (15b) is invalid (this symptom was pointed out by Frege, I believe). Third, it is not generally possible to substitute contingently coreferential expressions for each other within a propositional attitude; the inference in (15c) is invalid, since Ralph may never even have heard of Ortcutt (this symptom was pointed out by Quine). (I use # to indicate invalid inference.)

(15) a. Ralph believes that Susan is taller than she is.
b. Ralph believes that a goat walked into the room.
#Therefore, there is a goat such that Ralph believes it walked into the room.
c. Ralph believes that the man he saw on the beach is a spy.
Ortcutt is the man Ralph saw on the beach.
#Therefore, Ralph believes that Ortcutt is a spy.
(Similar judgments with all other verbs of situational attitude.)

The same symptoms occur in the infinitival complement of *intend*.[4]

(16) a. Ralph intends to give away more than he has.
b. Ralph intends to buy a goat.
#Therefore, there is a goat such that Ralph intends to buy it.
c. Ralph intends to shoot the man he saw on the beach.
Ortcutt is the man Ralph saw on the beach.
#Therefore, Ralph intends to shoot Ortcutt.
(Similar judgments with all other verbs of actional attitude.)

[3] Referential opacity has received copious discussion in the literature, starting with Russell 1905 and Quine 1956 and moving through an unmentionably large number of subsequent works. See Linsky 1971, Heny 1981, Searle 1983, and Jackendoff 1983, chap. 11, for representative references.

[4] Note that these symptoms do not appear with *all* infinitival complements, just with those that express actional attitudes. For instance, *manage to* is not a verb of actional attitude. If *managed* is substituted for *intends* in (16), (16a) becomes anomalous or at least sarcastic, and the inferences in (16b,c) go through.

That is, intentions, like beliefs, are individuated by "narrow content": the "referentially opaque" readings that lead to the judgments in (15)–(16) arise because the clause describes the situation or action in question (in part) from the point of view of the holder of the attitude. (The parallel between situational and actional attitudes is also pointed out by Searle (1983); for a more extended example, see note 14.)

Unfortunately, verbs of situational and actional attitude cannot be reliably distinguished by their syntax. It is true that *that*-clauses typically go with situational attitudes and infinitival clauses with actional attitudes, but this is not invariably the case. For instance, *wish* and *claim* can appear with an infinitival clause and nevertheless express a situational attitude.

(17) John wished/claimed...

$$\left\{\begin{array}{l}\text{to be shorter than Bob.}\\ \text{to have been born 10 years earlier.}\\ \text{to be descended from royalty.}\end{array}\right\}$$

More crucially for the present analysis, *intend* can occur with a subjunctive *that*-clause or a *for-to* infinitive, which need not be self-initiated, nor even express an action.

(18) a. John intended that Sue bring a cake.
b. John intends for Fred to be hit by a falling brick.

However, closer examination reveals that the *meaning* of (18a,b) does obey the constraint on *intend*, even though the syntax does not: the interpretation of (18a,b) is *coerced*[5] into a meaning in which John (the intender) acts to *bring about* the situation described by the complement; that is, the meaning actually does involve a self-initiated action. For example, (18a,b) can be fairly well paraphrased by (19a,b), in which *intend* is followed by a VP infinitival.

[5] When the interpretation of a sentence contains extra semantic material that is not contributed by any of the lexical items in the sentence, but that must be present in order for the sentence to be semantically well-formed, it is said to be *coerced* (Pustejovsky 1995). The clearest sort of example is due to Nunberg (1979).

(i) [One waitress to another:] The ham sandwich wants more coffee.

Since *want* requires an animate subject, the interpretation of *ham sandwich* in (i) is coerced into meaning 'person contextually associated with the ham sandwich'. For more discussion of coercion in general, see Jackendoff 1997a, chap. 3, and 2002a, sec. 12.2. For discussion of coercion with verbs like *intend*, see Culicover and Jackendoff 2005, chap. 12. For psycholinguistic evidence on the processing of sentences that involve coercion, see Piñango, Zurif, and Jackendoff 1999, Piñango and Zurif 2001.

(19) a. John intended to bring about that Sue bring a cake.
b. John intended to bring about that Fred be hit by a falling brick.

And the range of subjunctive *that*-clauses and *for-to* clauses that can occur with *intend* coincides with the range of things that the intender could bring about, as shown by comparing (18)–(19) with (20).

(20) a. *John intends that the sky be cloudy.
*John intends for Fred to be 10 years younger.
b. *John intends to bring it about that the sky be cloudy.
*John intends to bring it about that Fred is 10 years younger.[6]

Note that this *bring about* paraphrase is not very felicitous when the subordinate clause is already an action performed by the intender.

(21) ??John intended to bring about . . .

$$\left\{\begin{array}{l}\text{that he look at Sue.}\\ \text{that he scratch his nose.}\\ \text{that he prove Fermat's theorem.}\end{array}\right\}$$

It is also not felicitous when the verb expresses a situational attitude.

(22) a. John claimed that Sue brought a cake. ≠ John claimed to bring it about that Sue brought a cake.
b. John wished that Fred would be hit by a falling brick. ≠ John wished to bring it about that Fred would be hit by a falling brick.

This is typical of coercion: it applies only when the "simple" interpretation of the sentence is anomalous, and it "fixes up" the interpretation.

In short, the original semantic generalization stands: verbs of actional attitude require their complements to be interpreted as self-initiated actions. The apparent counterexamples are only violations in the syntax: coercion inserts extra semantic material that permits the offending complements to be interpreted in accordance with the constraint. We will encounter several more cases of coercion in this chapter.[7]

[6] One can imagine situations where these sentences are good: for example, John is painting the sky for a theater backdrop, or he is falsifying Fred's passport or choosing an actor to play the part of Fred. On such scenarios, both (20a) and (20b) are all right, preserving the correlation.

[7] A further use of *intend* means something like 'intend to say' or 'intend to convey', as in *Fred intends (by that remark) that he hates linguistics*. The 'intend' component of this use also obeys the constraints on ordinary *intend*.

Despite the irregularities just observed in what kind of complement goes with what kind of verb, there is an interesting correlation that brings out the relation between situational and actional complements. It turns out that quite a few verbs in English express a situational attitude when followed by a *that*-clause, but an actional attitude when followed by an infinitival.[8]

(23) a. John persuaded/convinced Bill...
that the sky was green.
that he has a big nose.
that Fermat's theorem was provable.
that Sue would bring a cake to the party.
(situational)
b. John persuaded/convinced Bill...
to look at Sue.
to scratch his nose.
to prove Fermat's theorem.
*to be shorter than Bob.
*to have been born 10 years earlier.
*to be descended from royalty.
(actional)
c. John agreed that he was born 10 years before Bill. (situational)
d. John agreed to look at Sue/*to have been born 10 years before Bill. (actional)
e. John swore/decided that he was born 10 years before Bill. (situational)
f. John swore/decided to look at Sue/*to be born 10 years before Bill. (actional)
g. It never occurred to John that he was descended from royalty. (situational)
h. It never occurred to John to look at Sue/*to be descended from royalty. (actional)

There are two possible ways one could account for the fact that so many different verbs show the same alternation between situational and actional attitudes. One approach would be to say that all the verbs hap-

[8] Some of the verbs in (23), such as *agree*, also allow subjunctive *that*-clauses, with a coerced action interpretation; some, such as *persuade*, do not. The verbs *remember* and *forget* also take *that*-clauses and infinitivals with differing interpretations, but they are more complex than the cases in (21). *John remembered/forgot that he should go* is factive. *John remembered/forgot to go* presupposes that John *should go* or that John is *supposed to go*. For some details, see Culicover and Jackendoff 2005, chap. 12.

pen to be ambiguous—basically that the facts in (23) are a strange coincidence. A more interesting approach is to say that the verbs express the very same attitude in either case and that the difference lies only in whether the attitude is taken toward a situation or an action. I will explore this second approach.[9]

This hypothesis has a striking consequence. When the verb *decide* is used with a *that*-clause, it means roughly 'come to believe' (e.g. *John decided that it was cloudy* = 'John came to believe that it was cloudy'). When it is used with an infinitival, it means roughly 'come to intend' (*John decided to leave* = 'John came to intend to leave'). Similarly, (23a) can be paraphrased as 'John caused Bill to come to believe that...', and (23b) as 'John caused Bill to come to intend to...'.[10] According to our hypothesis, *decide* and *convince* express the same attitude in both cases. This leads us to conclude that *believe* and *intend* also express exactly the same attitude, in one case directed toward a situation (or proposition) and in the other toward an action.

One way of checking this hypothesis is to see if there are languages in which *believe* and *intend* translate into the same word—just as their inchoatives ('come to X') and causatives are expressed by the same word in English. Preliminary investigation suggests that there are such languages. Ken Hale, Moira Yip, and Virginia Yip have independently observed (pers. comm.) that in some contexts in Mandarin the word *xiang* appears to take either meaning: *wô xiâng tā jīntiān bú hùi lái* 'I think/believe he'll be coming today' versus *wô xiâng shì shi* 'I intend/would like to try'. Sylvain Bromberger has observed (pers. comm.) that the French verb *penser*, normally translated as 'think/believe', can also be used in contexts like *Je pense partir* 'I intend to leave'.

Even if *decide* is the same attitude when applied to a situation and an action, the meanings of the two cases are not identical. For instance, the actional attitude *John decided to stop smoking* is very close in meaning to the situational attitude *John decided that he should stop smoking* (actually an attitude toward a norm). But as pointed out in Searle 1983, Jackendoff

[9] Bratman (1999) attributes a similar view to Castañeda (1975). However, Castañeda characterizes the arguments of *believe* and *intend* as propositions and intentions rather than situations and actions, respectively, and the supporting linguistic evidence offered here is absent.

[10] On the validity of paraphrase as a test for semantic structure, in particular with causatives, and in particular answering Fodor's (1970) objections against such tests, see Jackendoff 1990, sec. 7.8; 2002a, sec. 11.2.

1985, and Bratman 1987, they are not synonymous: either can be true without the other, as seen in (24a,b). Compare these with (24c), whose verb *claim* expresses a situational attitude with both *that*-clauses and infinitivals. Here the counterpart *is* a contradiction (marked with #).

(24) a. Although John decided to stop smoking, he didn't decide that he SHOULD stop smoking.
b. Although John decided that he SHOULD stop smoking, he didn't actually decide TO stop smoking.
c. #Although John claimed to have stopped smoking, he didn't claim that he stopped smoking.

Under the present hypothesis, we correctly predict parallel results for *believe* and *intend*: (25) shows the nonsynonymy between an intention and a nearly equivalent belief about a future situation.

(25) a. Although John intends to stop smoking, he doesn't believe that he WILL stop smoking.
b. Although John believes that he WILL stop smoking, he doesn't actually INTEND to stop smoking.

8.4 The Folk Metaphysics of Actional Attitudes

Let's consolidate what we have found so far. Animate actions are a special subclass of situations. Actional *attitudes* are very much parallel to situational attitudes, in particular also creating opaque contexts. However, they are not a subclass of situational attitudes, since they are semantically distinct. There are three symptoms of these differences. First, the action must be self-initiated; that is, the Actor of the action must be identical to the holder of the attitude. Second, the action must not be temporally previous to the time at which the attitude is held. Third, when the same verb can express both a situational and an actional attitude, the two sentences are not synonymous. Let's get a feel for why all this should be the case.

Impressionistically, there has to be a way for the mind to convert conceptualized actions into actual executed actions—the transcendental act of will. But you can't just execute actions the moment they are conceptualized: that would result in your acting entirely on impulse. This would make it impossible, for instance, to work up a *sequence* of actions before carrying it out; one would willy-nilly carry out whichever piece of the sequence one happened to think of first. And it would be impossible to hold an instruction in mind, then choose when to carry it out. Thus, if there is to be any complexity in planning and behavior, it must be possible to conceptualize an action with a time attached to it that is different from the present—or with an unspecified time—and to just store this action plan in memory for later use.

We use verbs of actional attitude to describe our impressions of manipulating our own conceptualized actions. This immediately explains the observed restrictions on the actions over which actional attitudes can be held. First, such actions must be self-initiated because these are the only kinds of actions one can conceivably execute. You can't move someone else's body through an act of will alone. Second, actional attitudes must be nonpast-directed, because future times are the only times one can usefully attach to actions whose execution one is contemplating.

Whether or not this impressionistic account has anything to do with the neuroscience, it does a rather nice job with the *folk* theory of actional attitudes, which derives from what it feels like to contemplate and perform a voluntary action. Thus the verbs of actional attitude can be thought of as expressing our *conceptualization* of the manipulation of conceptualized actions prior to their execution. On this conceptualization, a voluntary action is one that arises from a conceptualized action.

8.5 The Conceptual Structure of *Believe* and *Intend*

8.5.1 States and Events versus Actions

I have been putting off formalization as long as possible, but the time has now come to get on with it. The first order of business is creating a formal difference between situations and actions, such that, on one hand, actions are a special subclass of situations but on the other hand, actional attitudes can be differentiated from situational attitudes.

In Conceptual Semantics, conceptualized entities are classified most crudely in terms of their *ontological categories*. Among the prominent ontological categories are *Object*,[11] *Situation*, *Place*, *Property*, *Amount*, and *Time*. In the notation, these appear as labels on the brackets that enclose the entity's descriptive content (these were omitted in the exposition of the notation in section 6.1, though *Property* appeared in chapter 7).

[11] Actually, *Object* is part of a more general class that also includes substances and assemblages of objects. In Jackendoff 1991, 1996c, this larger class is called *Material* and differentiated into its subclasses by a set of features. The same features also classify events into closed events (accomplishments), processes, and aggregate events (such as a lot of frogs jumping at once or in succession).

For instance, (26) repeats the conceptual structure of the sentence in figure 1.1, *The little star's beside a big star*.

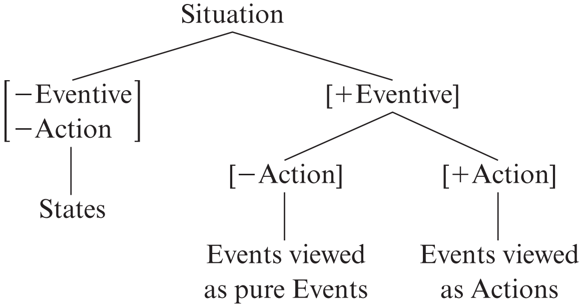
(26)
$$[_{\text{Situation}}\ \text{PRES}\ [_{\text{State}} \begin{bmatrix} & \text{STAR} \\ & [_{\text{Property}}\ \text{LITTLE}] \\ _{\text{Object}} & \text{DEF} \end{bmatrix} \text{BE}\ [_{\text{Place}}\ \text{BESIDE} \begin{bmatrix} & \text{STAR} \\ & [_{\text{Property}}\ \text{BIG}] \\ _{\text{Object}} & \text{INDEF} \end{bmatrix}]]$$

Here the outermost Situation consists of some State obtaining in the present; the State consists of some Object being at some Place; the two Objects are both stars, one with the Property 'little' and the other with the Property 'big'.

One way to capture the difference between situations and actions would be to introduce a new ontological category Action, distinct from Situation. However, there is something suspect about making such a distinction, since, as seen in section 8.2, Actions appear to be a particular kind of Situation. The usual account under such circumstances is to say that the related categories differ by a feature. For example, since Events and States are both kinds of Situations, we can categorize them as Situations that are *[+Eventive]* and *[−Eventive]*, respectively.

In turn, Actions are a subset of Events. For reasons to become clear, I would like to treat this distinction by saying that certain Events have two possible conceptualizations: one as a pure Event (something happening) and one as an Action (something an Actor is doing). Under the latter conceptualization, they have the feature *[+Action]*; under the former, they fall in with all other Situations as *[−Action]*. The resulting taxonomy of features is shown in (27).

(27)



In terms of this taxonomy, the categorization of the attitudes is simple: verbs of situational attitude such as *believe* apply to [Situation, −Action] constituents, and verbs of actional attitude such as *intend* apply to [Situation, +Action] constituents. Moreover, we can say that ambidextrous attitude verbs such as *decide* apply to any sort of [Situation] constituent. When the situation in question is [−Action], *decide* functions as a verb of situational attitude; when the situation is [+Action], *decide* is a verb of actional attitude.

What distinguishes those Events that can be construed as Actions from those that cannot? Basically, the criterion is that the subject can be construed as *doing* something, as shown by the *do* test (*What John did was*...) illustrated in (7). In the typical cases that pass this test, the sentence is active rather than passive, and the subject is either in motion or causing another Event. (28), based on the examples in (7), illustrates the distinction.

(28) a. *Construable as Actions*
Zukerman performed *Harold in Italy*. (subject causes music to be produced)
The US invaded Iraq. (subject is in motion (among other things))
b. *Non-Action Events*
Bill received a letter. (letter, not Bill, is in motion)
Fred was hit by a falling brick. (passive sentence; brick is in motion)

In other words, the sentences in (28a) can be understood as either [+Action] or [−Action]; those in (28b), though also Events, can be understood only as [−Action].

The reason for the treatment of the feature [±Action] now becomes evident. We would like to say that Zukerman's performing counts as a potential Action in the actional attitude *Zukerman decided to perform*, but it counts as a plain Event in the situational attitude *Zukerman decided that he would perform*. That is, all the descriptive content of the Event is the same in both cases: it is just that the Event is being regarded differently as a whole. (There may be some other tricky way to do this, but I can't think of one that makes everything else work so nicely. You have to choose your battles.)

The notion of Actor was worked out formally in section 6.2. It does not correspond to any of the standard thematic roles such as agent and theme, since either an agent or a theme can function as Actor. (29) reviews examples from section 6.2.

(29) a. Bill threw the ball. (*Bill* is Actor: *What Bill did was throw the*
agent theme *ball.*)
b. The car hit the tree. (*the car* is Actor: *What the car did was hit*
theme goal *the tree.*)

Section 6.2 encoded the role Actor as the first argument of the macrorole tier function AFF. Thus those situations that can be construed as [+Action] have the macrorole tier X AFF ⟨Y⟩ (intuitively, 'X affects the situation'). X simultaneously holds a role in the thematic tier, for instance agent or theme.

8.5.2 Formalizing Situational and Actional Attitudes

Next let us turn to the attitudes: believing, intending, considering, imagining, and so on. Let's notate the general predicate under which they all fall as *ATTITUDE.* Thus BELIEVE and INTEND are special cases of ATTITUDE, the same way POODLE is a special case of DOG and DOG is a special case of ANIMAL.

Holding an attitude is a State. This state has two arguments. The first is the individual X holding the attitude, which must be Animate (a special class of Object) and preferably (but not necessarily) a Person on the social plane of chapter 5. The second argument is the Situation or Action over which the attitude is held. So the overall form can be notated as (30). (The subscripts on the brackets denote ontological category again.)

(30) $[_{\text{Situation}, -\text{Eventive}}\ [_{\text{Animate/Person}}\ \text{X}]\ \text{ATTITUDE}\ [_{\text{Situation}}\ \text{Y}]]$

Situational attitudes such as BELIEVE add the further restriction that the argument Y is [−Action]; actional attitudes such as INTEND require this argument to be [+Action]. Thus amplified, the form for attitudes becomes (31a,b).

(31) a. *Situational attitude*
$[_{\text{Situation}, -\text{Eventive}}\ [_{\text{Animate/Person}}\ \text{X}]\ \text{ATTITUDE}\ [_{\text{Situation}, -\text{Action}}\ \text{Y}]]$
b. *Actional attitude*
$[_{\text{Situation}, -\text{Eventive}}\ [_{\text{Animate/Person}}\ \text{X}]\ \text{ATTITUDE}\ [_{\text{Situation}, +\text{Action}}\ \text{Y}]]$

Next we have to incorporate the special restrictions on actional attitudes. These can be stated as further restrictions on (31b). First consider the restriction that the Action argument must have an Actor identical to the holder of the attitude. Because the argument is [+Action], it must by stipulation contain the macrorole tier X AFF. Thus the constraint we want is that the Actor role X in this tier must be identical to the first argument of the attitude. This constraint can be stated by *binding* the Actor role to the holder of the attitude. The notation for binding (Jackendoff 1990) places a Greek letter in the bound role and a corresponding letter as a superscript on the position to which it is bound.[12] Applying this notation to actional attitudes, we get a structure like (32). (For mnemonic convenience, I will substitute the notation *X ACT* for *X AFF* in the rest of this chapter.)

(32) $[_{\text{Situation, −Eventive}}\ \text{X}^{\alpha}\ \text{ATTITUDE}\ [_{\text{Situation, +Action}}\ \alpha\ \text{ACT}]]$

The second constraint on actional attitudes is that the time of the Action must not precede the time of the attitude. To incorporate this into the formalism, it is necessary to introduce a Time constituent. Time is not part of the argument structure of the verb; rather, it is signaled by a combination of time adverbials and tense. Accordingly, I will notate the Time of a situation connected by a semicolon to the function-argument structure of the situation, as in (33).

(33) $[_{\text{Situation}}\ \text{S};\ [_{\text{Time}}\ \text{T}]]$
‘Situation S obtains at time T.’

An intention involves two times, the time at which the attitude is held and the time of the contemplated action.

(34) $[_{\text{Situation}}\ \text{X}^{\alpha}\ \text{ATTITUDE}\ [_{\text{Situation, +Action}}\ \alpha\ \text{ACT};\ \text{T}_2];\ \text{T}_1]$
‘At T_1, X has an actional attitude toward acting at T_2.’

Again, the condition for future-directedness is that T_2, the time of the contemplated action, is not earlier than T_1, the time at which the attitude is held. I will express this with the same sort of binding notation used in (32), this time on the Time constituent.

(35) $[_{\text{Situation}}\ \text{X}^{\alpha}\ \text{ATTITUDE}\ [_{\text{Situation, +Action}}\ \alpha\ \text{ACT};\ [_{\text{Time}}\ \text{T}_2 \geq \beta]];\ \text{T}_1^{\beta}]$
‘At T_1, X has an actional attitude toward acting at a time subsequent to T_1.’

(35) can be taken as a well-formedness condition (or an axiom) for a conceptual structure to count as a coherent actional attitude. Whenever an attitude is formulated whose second argument is [+Action], all the

[12] Reflexive pronouns are a typical case of such binding. If *Bob washed Harry* is notated as (i), then *Bob washed himself* can be notated as (ii), where the α serving as the second argument of WASH is bound to the first argument. See Jackendoff 1990, Culicover and Jackendoff 2005, chaps. 6, 10, 11, 12.

(i) $[_{\text{Situation, +Eventive}}\ \text{BOB WASH HARRY}]$

(ii) $[_{\text{Situation, +Eventive}}\ \text{BOB}^{\alpha}\ \text{WASH}\ \alpha]$

structure in (35) is automatically supplied—a sort of coercion. This structure, which produces a fundamental asymmetry between situational and actional attitudes, is a consequence of the fact that an actional attitude must be toward something one can potentially perform oneself.

Now, looking specifically at *believe* and *intend*, the hypothesis from section 8.3 is that they express the very same attitude, differing only in that *believe* is directed at a non-Action and *intend* at an Action. The element they share might be expressed as 'commitment': to believe a situation is the case is to be committed to its existence, and to intend to do something is to be committed to doing it. Accordingly, I will call the attitude in question *COM* ('commitment'). Thus the conceptual structure assigned to *believe* is (36a), and that assigned to *intend* is (36b).[13]

(36) a. X believes that P.
$[\mathrm{X\ COM}\ [_{\mathrm{Situation},\,-\mathrm{Action}}\ \mathrm{P}]]$
b. X intends to act.
$[\mathrm{X}^{\alpha}\ \mathrm{COM}\ [_{\mathrm{Situation},\,+\mathrm{Action}}\ \alpha\ \mathrm{ACT};\ [_{\mathrm{Time}}\ \mathrm{T}_2 \geq \beta]];\ \mathrm{T}_1^{\beta}]$

As noted earlier, the verb *decide* is the inchoative of both: it can mean either 'come to believe' or 'come to intend'. And *convince* is the causative of both: it can mean either 'cause to come to believe' or 'cause to come to intend'. Thus these verbs can be formalized as (37a,b) (*INCH* may be read 'it comes about that').

(37) a. X decides that P/to P.
$[_{\mathrm{Situation},\,+\mathrm{Eventive}}\ \mathrm{INCH}\ [\mathrm{X\ COM}\ [_{\mathrm{Situation}}\ \mathrm{P}]]]$
b. Y convinces X that P/to P.
$[_{\mathrm{Situation},\,+\mathrm{Eventive}}\ \mathrm{Y\ CAUSE}\ [\mathrm{INCH}\ [\mathrm{X\ COM}\ [_{\mathrm{Situation}}\ \mathrm{P}]]]]$

Just in case P is [+Action], template (35) comes into play and imposes the appropriate restrictions for an actional attitude, so that the result is (38).

(38) a. X decides to act.
$[_{\mathrm{Situation},\,+\mathrm{Eventive}}\ \mathrm{INCH}\ [\mathrm{X\ COM}$
$[_{\mathrm{Situation},\,+\mathrm{Action}}\ \alpha\ \mathrm{ACT};\ [_{\mathrm{Time}}\ \mathrm{T}_2 \geq \beta]];\ \mathrm{T}_1^{\beta}]]$
b. Y convinces X to P.
$[_{\mathrm{Situation},\,+\mathrm{Eventive}}\ \mathrm{Y\ CAUSE}\ [\mathrm{INCH}\ [\mathrm{X\ COM}$
$[_{\mathrm{Situation},\,+\mathrm{Action}}\ \alpha\ \mathrm{ACT};\ [_{\mathrm{Time}}\ \mathrm{T}_2 \geq \beta]];\ \mathrm{T}_1^{\beta}]]]$

[13] A third kind of commitment might be commitment to a norm (an N-value in the sense of chapter 9). For example, to *believe **in** doing X* is not necessarily to intend to do it, but it entails (a) a belief that doing X is normatively good, plus (b) a commitment to adhere to this norm.

To be clear about the formalism, let's compare the nearly equivalent beliefs and intentions in (39). The only substantive difference between them is that the belief treats *he will stop smoking* as a pure Event that John is thinking about, while the intention treats the same Situation as an Action that John is contemplating. The Time of the main clause, being present tense, is NOW. The future tense in the subordinate clause of (39) happens to be later than NOW, but the Time of the intended Action is *bound* to being later than NOW.

(39) a. John believes that he will stop smoking.

$$[\text{JOHN COM} \\ \left[\begin{array}{ll} & \text{JOHN}^{\alpha}\ \text{STOP SMOKING; T} > \text{NOW} \\ {}_{\text{Situation, }-\text{Action}} & \alpha\ \text{ACT} \end{array}\right]; \text{NOW}]$$

b. John intends to stop smoking.

$$[\text{JOHN}^{\alpha}\ \text{COM} \\ \left[\begin{array}{ll} & \alpha\ \text{STOP SMOKING; T} \geq \beta \\ {}_{\text{Situation, }+\text{Action}} & \alpha\ \text{ACT} \end{array}\right]; \text{NOW}^{\beta}]$$

8.5.3 COM as a Conceptualization of a Valuation Feature: Theory of Mind Yet Again

Backing off from the formalism for a moment, what sort of concept is COM? Recall again the discussion of section 3.3. There I proposed that the character of awareness is determined in part by *valuation features*, which give percepts their "feel." Among the features proposed was a feature [±committed], which applies to percepts that one senses as meaningful. I proposed further that this feature has three subcases:

- [+committed: valence+] pertains to percepts that one regards as "real" —percepts one "believes in."
- [+committed: valence−] pertains to percepts that one regards as "unreal"—things such as unbidden visions.
- [−committed] pertains to percepts to whose reality one has no commitment one way or the other, such as images one is entertaining.

In this light, COM appears to be the conceptualization of the "feel" of positive commitment in awareness. However, as in the case of EXP, discussed in sections 6.4 and 7.7, this predicate can be used not just to refer to one's own sense of commitment: it is also the means by which we attribute this valuation to others. In other words, like EXP, the predicate COM is an essential element of theory of mind. An organism that lacks this predicate can still feel committed to the reality of percepts (or have an experience of reality), because the valuation features are present in any event. But such an organism cannot attribute such percepts to others, and it cannot question its own commitments, because valuation features are not themselves elements of conceptual structure and do not enter into rules of inference. These cognitive tasks require the predicate COM, which is as it were the cognitive simulation of the valuation feature [+committed: valence+].

Certain other situational and actional attitudes are conceptualizations of other combinations of valuation features. (40) reviews some of the possibilities discussed in section 3.3.4.

(40) a. *Propositional attitudes*
[+committed: valence+]: believing that something is the case
[+committed: valence−]: disbelieving/doubting that something is the case
[−committed]: entertaining a proposition
[−committed; +affective: valence+]: desiring/wanting that something be the case
[−committed; +affective: valence−]: dreading/fearing that something is the case
b. *Actional attitudes*
[+committed: valence+]: intending to do something
[+committed: valence−]: avoiding doing something
[−committed]: considering doing something
[−committed; +affective: valence+]: desiring/wanting to do something
[−committed; +affective: valence−]: dreading/fearing to do something

I leave it for further research to determine how the whole family of attitudes is to be formally elaborated in terms of sister concepts to COM.

8.6 Doing Something Intentionally, the Volitionality of Action, and Imperatives

With the formalization of intending in (36b), it becomes possible to formulate a whole range of notions centered around intentions.

8.6.1 Doing Something Intentionally

Let us consider what it means to do an action intentionally, a preoccupation of Searle (1983) and Bratman (1987). Bratman points out that although one may perform a particular action intentionally, one does not necessarily intend the consequences. For example, he observes that if I intentionally run home in the rain, it does not mean I intentionally get my shoes wet. Similarly, Searle discusses a grisly scenario in which he sets out in his car to kill his uncle, and on the way he is so agitated that in the fog he accidentally strikes and kills a pedestrian—who happens, unbeknownst to him, to be the very uncle he had set out to kill. Now it is clear that Searle has not killed the pedestrian intentionally: the act of striking and killing this pedestrian does *not* fulfill Searle's intention. In fact, he may well continue grimly on his way to his uncle's house, knife firmly in hand, intention firmly in mind.[14]

Searle's analysis of these scenarios is that one can perform an action intentionally only by performing it *with the intention that it fulfill an intention to perform that action.* But how is this more complex intention fulfilled? It looks as though we are heading for an infinite regress. A possible way out involves a tricky use of the binding notation, shown in (41).

(41) X intentionally performs action Y.

$$\left[\begin{array}{l} \text{Y} \\ \text{X}^{\alpha}\ \text{ACT} \\ [\text{FROM } [\alpha \text{ COM } [_{\text{Situation}, +\text{Action}}\ \beta]]] \end{array}\right]^{\beta}$$

[14] Searle's hit-and-run situation provides a good illustration of referential opacity in intentions, of the sort discussed in section 8.3. Under his scenario, both sentences in (i) have true readings.

(i) a. Searle did not intend to kill the unknown pedestrian.
  b. Searle intended to kill the unknown pedestrian.

The reason for this curious situation is that there is a systematic ambiguity in all the terms within the description of an attitude. They may record the way the holder of the attitude describes the contemplated event to him- or herself (the *opaque* description, a.k.a. a description of the *narrow content* of the attitude). Or they may record the speaker's description of the individuals, situations, or properties to which the attitude pertains, independently of the way the holder of the attitude may happen to describe them to him- or herself (the *transparent* description, a.k.a. a description of the *broad content* of the attitude). Assuming the narrow content of Searle's intention is expressed by 'I intend to kill my uncle', (ia) is true on the opaque reading of *the unknown pedestrian.* But because, unbeknownst to Searle, the person *he* describes as 'my uncle' happens also to be describable as 'the unknown pedestrian', (ib) is true on the transparent reading of *the unknown pedestrian.* (Of course, (ia) is false on the transparent reading and (ib) is false on the opaque reading.)

What does this say? An English paraphrase that unpacks it pretty well is 'X does Y out of an intention to do so'. The upper line encodes the thematic content of the action Y, whatever it may be. The second line is the macrorole tier, in which X is the Actor. The crucial part on the bottom line is a modifier. Its main function, notated as *FROM*, marks its argument as a cause (as in *They died **from** hunger*).[15] The argument of FROM is the situation of α's intending some Action β. In turn, because of the way β is bound, α's intention is toward the Action itself, *complete with intention*. That is, the intention is in part self-referential (Searle's term is "causally self-referential"). The fact that β must satisfy the constraints on the actional argument of *intend* guarantees that the main function Y, which is bound to β, is also an Action.

Let us now return to Bratman's and Searle's scenarios. Getting my shoes wet does not count as intentional, because it is not the act to which the intention is directed—that is, the commitment is not to *that very act*. Similarly, the hit-and-run death of Searle's uncle does not count as an intentional killing, because the running down of the unknown pedestrian is not the act to which Searle is committed. This shows up in the formalism as a failure of the intended action (β in (41)) to be able to bind the action as a whole. The action of getting my shoes wet cannot be bound by the intended action of running home; only the action of running home can. And the narrow content of the intention (see note 14) is important: Searle's action of killing (the person he identifies as) the unknown pedestrian cannot be bound by the intended action of killing (the person he identifies as) his uncle.[16]

[15] FROM is proposed in Jackendoff 1990, sec. 5.4. Intuitively, its argument is a Situation that brings about the Event that it modifies (e.g. *Hunger caused him to die*); that is, it is an inverse of CAUSE. Thus a possible unpacking of FROM Z is as (i), 'the property of being caused by Z', in which case the last line of (41) comes out as (ii).

(i) $[_{\text{Property}}\ \lambda x(\text{Z CAUSE x})]$

(ii) $[_{\text{Property}}\ \lambda x([\alpha \text{ COM } [_{\text{Situation}, +\text{Action}}\ \beta]] \text{ CAUSE x})]$

[16] Linguists will be concerned with what semantic structure is stored as the meaning of the adverb *intentionally*. Actually, (41) itself will do the trick, if Y and X^{α} ACT are treated as contextual features. In other words, when the adverb *intentionally* is added to the phonology and syntax of a sentence whose conceptual structure is otherwise Y plus X^{α} ACT, the modifier and binding in (41) are added to the conceptual structure.

8.6.2 The Intentional Stance

An important part of the folk theory of mind is a default assumption that actions are performed intentionally; that is, whenever possible, intentions are assumed to lie behind actions. That's how we manage to reason about other minds without being able to observe them. Of course, we often make mistakes, attributing intention to people for actions they do not (claim to) intend. But this assumption works a lot of the time. In addition, it is the reason that we have a tendency to anthropomorphize inanimate objects that initiate action, such as wind, clouds, and especially computers, saying they "want" to do whatever they do.

Formally, this assumption, part of Dennett's (1987) *intentional stance*, can be stated as a default (or defeasible) inference rule of the form (42).

(42) *The intentional stance*

$$[\text{X ACT}] \Rightarrow_{\text{default}} \begin{bmatrix} \text{X}^{\alpha}\ \text{ACT} \\ [[\text{FROM } [\alpha \text{ COM } [_{\text{Situation}, +\text{Action}}\ \beta]]] \end{bmatrix}^{\beta}$$

That is, unless there is evidence otherwise, we assume that any action is intentional.

Rule (42) doesn't just relate to the philosophical/psychological issue of attribution of intention. It also permits an interesting result in linguistic semantics. It has always been noticed that verbs like *roll* and *slide* are ambiguous with respect to whether their subjects are volitional or not: John's rolling or sliding down the hill may be a result of his having decided to do it, or it may be a result of his having been pushed. When the issue has been raised, the standard assumption (mine anyway) has been that all these verbs are polysemous—that they have an optional feature of volitionality in their lexical entries. But why should this be true of *all* these verbs (including even psychological verbs like *surprise*; see chapter 7), rather than just some of them? An optional lexical feature leaves this generalization unexplained.

Suppose, though, that among the general principles of sentence interpretation is the rule (42) (be it a rule of pragmatics, a Gricean implicature, a principle of structural meaning in the sense of Gleitman et al. 1996, or a principle of coercion in the sense of Jackendoff 1997a). Then *any* action verb with an animate subject will automatically present the possibility of a volitional interpretation. Thus the feature of volitionality will not have to be included in the lexical entries of these verbs at all, which simplifies the lexicon considerably. (However, verbs like *murder* that *require* a volitional subject will still include some volitional predicate in their lexical entries.) On this analysis, an important part of the semantic content of a sentence like *John rolled down the hill* arises not from its words but from principle (42).

8.6.3 Imperatives

For what I find a rather unexpected result, consider the relation between declarative and imperative sentences.

(43) a. Bill ate the pizza.
b. Eat the pizza.

One of the standard tests for whether a VP expresses a voluntary action is whether it can be used as an imperative. The reason is that the well-known selectional restrictions on imperatives are identical to those for intended actions. Imperatives require the understood subject YOU to be a volitional Actor. Hence imperatives based on statives (44a) and non-self-controllable events (44b) are unacceptable.

(44) a. *Be descended from royalty.
b. *Grow taller.

Where at all possible, such unacceptable cases are coerced into causative readings.

(45) a. Be quiet. = Make yourself quiet.
b. Be examined by a doctor. = Get yourself examined by a doctor.

An imperative, moreover, has to be nonpast-directed.

(46) Leave the room now/in five minutes/*five minutes ago.

This parallelism should not be a coincidence. Accordingly, suppose the conceptual structure corresponding to imperative force is simply (47).

(47) Do such-and-such.
$[_{\text{Situation}, +\text{Action}}\ \alpha\ \text{ACT}]$

How will imperative force follow from this? Think about how a sentence acquires its illocutionary force. Just about every theory of illocutionary force supposes that conventional communication pastes some additional material around the content of the sentence; theories differ in precisely what this additional material is. For instance, (48) gives an informal account of the extra information pasted around a declarative sentence (adapting for convenience the style of Wierzbicka 1987).

(48) For a declarative sentence S with conceptual structure $[_{\text{Situation}}\ \text{P}]$, the illocutionary force is:

I am saying S to you
out of the intention to cause you to come to believe $[_{\text{Situation}}\ \text{P}]$.

For present purposes, the crucial part of this is *you believe*, which in the present formalization comes out as (49).

(49) ... out of the intention to cause $[\text{YOU COM } [_{\text{Situation, }-\text{Action}}\ \text{P}]]$

Suppose that we substitute into this formula, instead of a declarative sentence, an imperative sentence with the meaning (47). Then instead of (49) we get (50a). Since the argument of COM is [+Action], the template for actional attitudes, (35), is imposed, yielding the full form (50b).

(50) a. ... out of the intention to cause
$[\text{YOU COM } [_{\text{Situation, }+\text{Action}}\ \alpha \text{ ACT}]]$
b. ... out of the intention to cause
$[\text{YOU}^{\alpha} \text{ COM } [_{\text{Situation, }+\text{Action}}\ \alpha \text{ ACT; } [_{\text{Time}}\ \text{T}_2 \geq \beta]];\ \text{T}_1^{\beta}]$

Notice now that the formalized part of (50b) corresponds to *you intend to act*. So, replacing *you believe* in (48) by *you intend*, we get (51).

(51) For an imperative sentence S with conceptual structure
$[_{\text{Situation, }+\text{Action}}\ \text{A}]$, the illocutionary force is:
I am saying S to you
out of the intention to cause you to come to intend
$[_{\text{Situation, }+\text{Action}}\ \text{A}]$.

This is just about right for the force of an imperative. In particular, it abstracts away from whether the sentence is meant as a request or an order. In addition, because A becomes the object of an actional attitude, it is automatically restricted to the structure (35), including an Actor bound to YOU and a time not prior to the present (when the imperative is uttered), as seen in (50b). That is, this very simple account of the semantics of imperatives predicts the observed restrictions as a special case of the restrictions on actional attitudes.

Thus, if we assume that the conceptual structure of an imperative is (47), we can unify the descriptions of the illocutionary force of declarative and imperative sentences with no further ado. The more general description is (52).

(52) For a declarative or imperative sentence S with conceptual
structure $[_{\text{Situation, }\pm\text{Action}}\ \text{X}]$, the illocutionary force is:
I am saying S to you
out of the intention to cause you$^{\alpha}$ to come to $[\alpha \text{ COM X}]$.

8.7 Fulfilling versus Voiding an Intention; Purposes

Fulfillment of an intention has the curious effect of wiping out the intention. Roughly speaking, if I intend to move my finger (or buy a new car), and actually do move my finger (or buy a new car), I no longer have the intention to do so.

What sort of entity is this that "goes out of existence" through the occurrence of an event that it describes? Our beliefs do not go out of existence when we find out they are correct. On the other hand, obligations and wishes (some kinds, anyway) do go away when they are satisfied. Bodily sensations such as hunger, thirst, itches, and the need to urinate also go away when the relevant events for their satisfaction take place. And so do even the needs of inanimates, such as a house's need for a new roof.

The reason this behavior seems strange lies in the grammar of the words *intention*, *wish*, and so on. Although they are nouns and therefore ostensibly denote some sort of abstract object, I suggest this is a grammatical illusion: they actually denote states rather than objects. There is no real difference in meaning between *intending* and *having an intention*. The latter just couches the thought in terms of a so-called light verb construction, in which the real content of the predicate is not in the main verb *have* but in the nominal *intention*. This situation is paralleled by many other light verb constructions. (For one proposal on how the light verb construction comes to have this interpretation, see Culicover and Jackendoff 2005, sec. 6.5.1.)

(53) a. John has the intention/a wish to go to Texas. (= John intends/wishes to go to Texas)
b. John has a tendency to sneeze. (= John tends to sneeze)
c. John took a walk to the store. (= John walked to the store)
d. John put the blame on Bill for the accident. (= John blamed Bill for the accident)
e. John gave a performance of the *Waldstein*. (= John performed the *Waldstein*)

There are no entities *tendency*, *walk*, *blame*, and *performance* independent of the events of someone tending to do something, someone walking, someone blaming, and someone performing: it would be a mistake to reify tendencies, walks, blame, and performances as though they had independent status. Similarly, it is a mistake to reify intentions, wishes, and other situational and actional attitudes—*including beliefs*! We conclude that an intention "ceases to exist" when the state of intending ends; the "fulfillment of an intention" is the termination of a state of intending as a result of performing the intended action.

However, an intention can go out of existence (i.e. a state of intending can end) for reasons other than its fulfillment. Suppose Amy intends to feed the cats, but then discovers Beth has already done so. As a result of this discovery, Amy no longer has her intention—although she has not fulfilled it. I will call this more general situation the *voiding* of an intention.

What is going on in this case is that we understand Amy's intention as really having more content than just feeding the cats. Rather, she intends to feed the cats in order to achieve some implicit purpose (or goal), presumably that the cats shouldn't go hungry. This is of course a pragmatic inference on our part: she may be determined to feed the cats come hell or high water, whether they want to be fed or not. But it is more charitable to attribute to her by default some reasonable purpose. And intuitively, her intention is voided if she comes to believe that the purpose is satisfied. (Note that it is not enough that the purpose be satisfied pure and simple: Amy's intention does not go away when Beth feeds the cats, but only when Amy *learns* that Beth has fed the cats.)[17]

A purpose behind an intention can of course be overtly expressed, for example in (54).

(54) a. John intended to go home in order to see his mother.
b. John intended to buy a car in order to get to work more easily.

Again, if through some other means John gets to see his mother, or another way develops for him to get to work easily (say, he inherits a motorcycle or a new bus line goes into operation), the intention may be voided.

Intuitively, a purpose can be thought of as something one wants that motivates an intention to act.[18] I will formalize this as (55).

[17] For a trickier case, let us return to Searle's story from section 8.6.1. At the point at which he has killed the unknown pedestrian, his intention is intact. But somewhat later he gets arrested for his hit-and-run accident and learns the identity of his victim. At this point, his intention is presumably voided. This case falls under the present one, if we attribute to Searle a reasonable although somewhat redundant purpose behind his intention to kill his uncle: that his uncle be dead. Since he now knows that the purpose is satisfied, the intention terminates.

[18] This applies only to purposes attributed to beings that can have intentions. Purposes can also be attributed to artifacts (*A phone rings to let us know someone's calling*) and to nonsentient living things (*A tree has leaves to collect sunlight*). These have somewhat different structure; an ideal analysis would unify them all as variations on a common theme. I won't do it here.

(55) X intends to do something in order for Z to come about.

$$\left[\begin{array}{l} \text{X}^{\alpha}\ \text{COM}\ [_{\text{Situation, +Action}}\ \alpha\ \text{ACT}] \\ [\text{FROM}\ [\alpha\ \text{WANT Z}]] \end{array}\right]$$

In (55), the purpose is treated as a FROM modifier to conform to its subordinate role in the syntax. The meaning is that X's desire for Z is what causes X to intend to act. (WANT is for present purposes treated as an unanalyzed situational attitude; this does not preclude further analysis as suggested in section 8.5.3.)

Now notice that wanting has the same characteristic as intending, namely that if a want is satisfied, it (usually) goes away. In (55), the want is what causes the intention. If the want ceases, the intention does too. In other words, the voiding of an intention by the independent fulfillment of its purpose follows from the representation in (55) (plus intuitively clear rules of inference that I will not try to formalize here[19]).

What about purposes with verbs other than *intend*, for instance *go* and *buy*?

(56) a. John went home in order to see his mother.
b. John bought a car in order to get to work quicker.

The clue to the analysis is that purposes can only pertain to voluntary actions. One cannot *grow taller in order to*... or *be descended from royalty in order to*..., and one cannot *buy a car today in order to get to work yesterday*. In other words, all purposes presuppose an intention. How does the intention get into the conceptual structure of (56a,b)? The simplest possibility is that it is supplied by the intentional stance—the rule of default entailment given in (42): "unless there is evidence to the contrary, assume an action is performed intentionally." By applying this rule, we construct an intention to which the purpose can be attached. (57) gives the resulting structure for (56a).

(57)

$$\left[\begin{array}{l} \text{JOHN}^{\alpha}\ \text{GO HOME} \\ \alpha\ \text{ACT} \\ [\text{FROM}\ \left[\begin{array}{l} \alpha\ \text{COM}\ [_{\text{Situation, +Action}}\ \beta] \\ [\text{FROM}\ [\alpha\ \text{WANT}\ [\alpha\ \text{SEE MOTHER}]]] \end{array}\right] \end{array}\right]^{\beta}$$

'John went home out of the intention to do so, where the purpose behind the intention was a desire to see his mother.'

[19] If, as suggested in note 15, [FROM Z] is an abbreviation for λx[Z CAUSE x], the rule of inference in question is fairly straightforward: if a cause doesn't take place, the effect usually doesn't either.

Let us inspect the inferences that follow from (57). First, we understand that John has gone home. Second, because his going home fulfills the intention to go home, he no longer intends to do so. Third, (57) does not assert that John saw his mother—only that he wanted to. Therefore we do not know if John's desire to see his mother has been fulfilled and thereby terminates. These inferences seem just right for (56a).

Finally, let us return to imperative sentences. There is a form of imperative that is used not to request or order, but to give instructions or advice.

(58) a. To make more money, work harder.
b. To bake a cake, take 3 cups flour, . . . , and put it in the oven for an hour.

Significantly, these sentences contain purpose clauses. Can we explain why these sentences have this force?

Let us elaborate a little further on the treatment of the illocutionary force of imperatives outlined in the previous section. Suppose that adding a purpose clause to an imperative adds it to the intention that the speaker wishes the hearer to adopt. Then the conceptual structure of an imperative with a purpose clause (including illocutionary force) comes out like (59).

(59) Act in order to Z.
I am saying S to you out of the intention to cause you$^{\alpha}$ to come to

$$\begin{bmatrix} \alpha \text{ COM } [\alpha \text{ ACT}] \\ [\text{FROM } [\alpha \text{ WANT } [\alpha \text{ Z}]]] \end{bmatrix}.$$

This is pretty close to the right force. For if the hearer does not want Z, and Z is what causes the hearer's intention to perform the contemplated action, then the intention terminates too. On the other hand, if the hearer does want Z, then the intention appears. That's what instructions and advice are: if you want Z, then do such-and-such.

Now recall the earlier example of Amy intending to feed the cats, with an implicit purpose whose satisfaction could void the intention—without her performing the action. Similar cases occur with imperatives.

(60) a. Shake well.
b. Do not use near fire or flame.

These are not commands, but instructions: if you want this object to function properly, then do the following thing. So the more general use of imperatives for instructions follows from our analysis of the logic of implicit purposes.

An important point arises in these examples: a good deal of utterance meaning is *not* present in the meanings of the words in the sentence. This point has come up already in connection with the coercion of subjunctive *that*-clauses with *intend* and with the default constructional meaning responsible for attributing intention to animate subjects (the intentional stance). By now this should not be a controversial issue, given the flourishing inquiry into pragmatics and especially into coercion and cocomposition (Pustejovsky 1995; Jackendoff 1997a). Nevertheless, I wish to stress it as a corrective to the assumption, widespread in linguistics and philosophy and dating back at least to Frege, that all semantic content comes from word meanings, simply pasted together in accordance with the dictates of their syntactic arrangement.

8.8 Joint Intentions

Chapter 5 introduced the notion of joint intentions and jointly intended actions as an essential part of cooperative behavior, based on discussions by Gilbert (1989), Searle (1995), Clark (1996), and Bratman (1999). Let's now augment the formalism of this chapter so it can encompass joint intentions. This is especially important because jointly intended actions are essential to the treatment of exchange transactions such as buying, selling, trading, and other contracts, which will be discussed further in chapter 10.

Recall the intuition behind a joint action. If we are moving a couch together, the task is more than the sum of my moving one end and your moving the other. If we are playing a duet, the task is more than your playing your part and my playing my part. Rather, the task has to be conceptualized in terms of our carrying it out jointly—you doing your part and I doing my part, with proper temporal coordination if necessary. In order to initiate and coordinate a joint task, the participants must signal their intentions. The signals may be linguistic:

(61) A: Let's move the couch. (offer to engage in joint task)
B: OK. (establishment of presumptive joint intention)
A: Ready? OK, now! (coordinating signal)

Or the signals may be gestural: holding out a hand to shake (chapter 4) counts as an offer. Even subtle unconscious inflections of motion are important, such as sensing when to release a handshake or a hug.

I will only take up a small part of this here, namely the conceptualization of joint actions and joint intentions. (I won't take up the crucial part of *how* the participants arrive at a joint intention, which may involve bargaining and even perhaps coercion and threat.)[20] The idea is that instead of the Actor being an individual, it is a set. (62) shows an action with two individuals as co-Actors (a generalization to more participants is intuitively clear, and I won't go through the formal hoops necessary to state it).

(62) X and Y jointly perform action.

$$[\{X, Y\} \text{ ACT}]$$

We must next decompose the action into the parts performed by X and by Y. In order to represent this, we need to introduce the operator *COMPOSED-OF*. This operator is important not only for joint actions, but also for articulating the components or ingredients of objects and substances. For instance, (63) is an encoding for an expression like *chicken noodle soup* (see Jackendoff 1991).

(63) chicken noodle soup

$$\begin{bmatrix} \text{SOUP} \\ \text{COMPOSED-OF } \{\text{CHICKEN, NOODLES}\} \end{bmatrix}$$

Applying this operator to the encoding of a joint action, we get an expression like (64), where ACT_x and ACT_y are X's and Y's respective contributions to the joint activity.

(64) X and Y jointly perform action A, where X's part is ACT_x and Y's part is ACT_y.

$$\begin{bmatrix} \{X, Y\} \text{ ACT}_A \\ \text{COMPOSED-OF } \{X \text{ ACT}_x, Y \text{ ACT}_y\} \end{bmatrix}$$

A joint intention by X and Y to perform action A then can be formulated as (65).

(65) X and Y jointly intend to perform action A.

$$[\{X^{\beta}, Y^{\gamma}\}^{\alpha} \text{ COM} \begin{bmatrix} \alpha \text{ ACT}_A \\ \text{COMPOSED-OF } \{\beta \text{ ACT}_{\beta}, \gamma \text{ ACT}_{\gamma}\} \end{bmatrix}]$$

[20] Bratman (1999) is concerned to exclude from joint intention the case of joint tasks undertaken because of coercion on the part of one participant (e.g. *We're going to New York together, OK?* [holding a gun to addressee's head]). I am inclined to include these cases as jointly intended tasks and to treat the coercion as part of the process by which the joint intention is formed, nefarious though this process may be.

To be careful about this: (65) is the conceptualization of a joint intention, as though X and Y do share a commitment to the same action. In order to have such a conceptualization of the situation, X and Y need not *really* share the same intention, and, as noted in chapter 5, this leaves plenty of room for misunderstanding and deception.

(65) now enables us to spell out more precisely the participants' commitments and their knowledge of the commitments of the others. There are two aspects to this. First, each participant intends to do his or her own part. But that isn't enough, because you may intend to do something and then not carry it out. So the second aspect is that each participant is obligated to the others to do his or her own part (Gilbert 1989; Bratman 1999). As we will see in chapter 11, this means that if X does not perform, then Y has a right to seek redress—perhaps an apology in a simple case, damages for breach of contract in a more elaborate one. (66) shows the formalization of obligation that will be developed in chapter 11.

(66) X is obligated to Y to perform action ACT.
X^{α} OB (α ACT, TO Y)

Given all these parts, we can now put together the following entailments for a joint intention:

(67) $[\{X^{\beta}, Y^{\gamma}\}^{\alpha}\ \text{COM} \begin{bmatrix} \alpha\ \text{ACT} \\ \text{COMPOSED-OF}\ \{\beta\ \text{ACT}_{\beta}, \gamma\ \text{ACT}_{\gamma}\} \end{bmatrix}] \Rightarrow$

a. X^{β} COM [β ACT_{β}]
b. Y^{γ} COM [γ ACT_{γ}]
c. X^{β} OB (β ACT_{β}, TO Y)
d. Y^{γ} OB (γ ACT_{γ}, TO X)

Because each of the participants can draw all these inferences—and knows that the other can as well—all of the expected conditions of mutual knowledge are met.

In general, joint actions need not be intentional: consider two chemicals interacting. But the joint actions we're interested in here are all intentional. Hence we need to combine the structure of a joint task (64) with the structure of intentional action. (68) is the complete structure.

(68) X and Y intentionally perform action A jointly.

$$\begin{bmatrix} \{X, Y\}\ \text{ACT}_{\text{A}} \\ \text{COMPOSED-OF}\ \{[\text{X ACT}_{\text{x}}], [\text{Y ACT}_{\text{y}}]\} \\ [\text{FROM}\ [\{X, Y\}\ \text{COM}\ \alpha]] \end{bmatrix}^{\alpha}$$

From the third line of (68), the joint intention to perform the joint task, the inferences in (67a,b) follow.

Note that certain uses of *with NP* and *together* will also induce the sense of a joint task.

(69) a. Sue baked a cake with Sally.
b. Sue and Sally baked a cake together.

Thus, just as *intentionally* introduces the modifier FROM [X COM α], so *together* and *with NP* can force a joint action structure. The technical details are beyond the scope of this book, though.

Given the analysis in the present chapter, it should go without saying that mutual belief falls out of the same notation, except that the right-hand argument of COM is a [−Action] constituent. The entailments for mutual belief parallel (67a,b): one can infer that both participants believe the proposition in question. However, there is no special part of the belief for which each participant is responsible, and there is no counterpart to (67c,d): there is nothing that the participants are expected to deliver for the benefit of the others.

(70) *Mutual belief and its entailments*
$[\{X, Y\}\ \text{COM}\ [_{\text{Situation, }-\text{Action}}\ \text{P}]] \Rightarrow$
a. [X COM P]
b. [Y COM P]

8.9 Conclusion

Summing up, we have analyzed *intend* as the function COM applied to an actional complement. The function COM brings *intend* into parallelism with *believe*, as motivated by the fact that they have parallel inchoatives (*decide*) and causatives (*persuade/convince*). The actional complement unifies *intend* with all the other actional attitudes and differentiates it from the situational attitudes. At the same time, all the attitudes can now be understood as predicates that encode aspects of theory of mind: they are conceptualizations of some of the valuation features that give experience its "feel" of reality or unreality.

The special properties of actional attitudes, codified in (35), account for the fact that the action toward which an actional attitude is held must be self-initiated and at a time not prior to the time of the attitude. As a consequence of these restrictions, the *that*-subjunctive complements of *intend* require coercion into a semantic form where the intender is understood to intend bringing about the event in question.

This treatment has led to a straightforward analysis of *doing Y intentionally*. The rule of defeasible inference in (42) codifies Dennett's intentional stance and enables us to eliminate the feature of optional volitionality from the lexical entries of a vast number of verbs. The analysis also enables us to give an account of purposes, their relation to intention, and the role of implicit purposes in voiding intentions, again showing the degree to which utterance meaning is richer than the meanings of the words alone. In addition, the analysis has led to a nice account of the illocutionary force of two sorts of imperatives. Finally, we have been able to encode joint intention and its entailments, which are crucial for the description of cooperative action. That ought to be enough consequences for one chapter.

Chapter 9
The Logic of Value

9.1 Overview

The end of chapter 5 alluded to the fundamental question of how a system of values can be grounded. An essential part of the answer has to be an understanding of what a system of values *is*, the issue to be taken up in this chapter.

As in chapters 6–8, the basic approach will not be to ask what values "are in the real world." They are not *anything* independently of the people who conceptualize them. Rather, in consonance with my overall goal of investigating mental structure, the question is how humans conceptualize values (especially unconsciously[1]) and how values play a role in governing people's judgments and behavior.

This question will be approached in part by using linguistic expressions of value as clues for the organization of conceptual structure. The results will be validated by the extent to which, by positing a relatively constrained set of conceptual building blocks, we can formally describe a rich variety of linguistic expressions and commonplace intuitions involving value. Through understanding the conceptual structures in which values are embedded, it is to be hoped that we will be in a better position to inquire into the evolutionary and cultural roots of systems of value.

As discussed in chapter 5, the approach parallels the investigation of human language: I am concerned with framing the conceptual system underlying all systems of human value in all cultures, no matter how laudable, peculiar, or repulsive they may seem to us. I will not aspire to decide what value system we *should* adopt, that is, to make value judgments over

[1] See Barth 1993 for discussion of how (in the present terms) conscious, verbalizable values may be quite different from implicit (unconscious) values, which can be detected through regularities of behavior.

value systems: I don't think that science can tell us what goals to adopt. Nevertheless, I concur with Greene (2003) and Doris and Stich (2005) that investigation of the present sort can be useful in addressing such issues: we may be able to learn that if one's goals are such-and-such, such-and-such a value system will or will not be useful in achieving them.

My overall hypothesis is that value is a conceptualized abstract property connected to (conceptualized) objects, persons, and actions. It is abstract because it is not directly perceptible. The value of an entity plays a role in various rules of inference that affect the way one reasons about the entity. Thus value serves as an intermediary in a system of logic—logic not in any standard propositional or formal sense, but in the sense of a conceptual/heuristic logic, as it were, a "folk logic." The rules of this logic are a sort of internal accounting system that helps connect many sorts of disparate objects, actions, and persons. Cultures vary in what values are conventionally assigned to what sorts of entities in what contexts, but, according to my hypothesis, the basic logic of value—that is, the internal accounting system—is to some degree universal: it establishes the terms in which judgments of value and inferences based on value are framed.

A system of values has three basic elements. First, in order for objects, persons, and actions to have values, there must be principles that give grounds for assigning values to them—that is, rules in which value appears in the consequent of the rule: "If such-and-such takes place, then such-and-such a value is assigned." These "input rules" are the entryways into the value system. Second, assigning values is of little use unless values have some effect on behavior. So there must also be principles that favor performing certain actions on the basis of values—that is, rules in which value appears in the antecedent of the rule: "If such-and-such an action has such-and-such a value, then do it." These are the "outputs" of the value system. Finally, in between input and output there may be many inferences that involve values in both the antecedent and the consequent—that is, reasoning internal to the value system.

An example of such an internal accounting system may make the idea clearer. Consider the conceptual status of points in a game (a case discussed in Searle 1995). Such-and-such a physical action in the game leads to the assignment of so-and-so many points to a player; this is the entry into the system. Internal to the system, points are totaled by addition as they are assigned; this is an inference that has no physical counterpart. The output of the system is the rule that says that the winner at the end is the player with the most points (or, depending on the game, the player with the fewest points, or the first player to reach a prescribed total). The points therefore serve as an inferential intermediary between the actions in the game and the outcome; without this intermediary, the outcome cannot be determined. The points themselves have no significance except within the context of the game (or the *frame* in the sense of chapter 5): it is senseless to say, out of the blue, "I have three points," as if it were like "I have three books."

Values are more complex than points for four reasons. First, a value has two dimensions: a *valence* (good (i.e. positive) or bad (i.e. negative)) and a *magnitude* (better or worse). Except in the case of monetary value, the magnitude is not a numerical quantity, but a relative quantity, perhaps measured by the basic mammalian magnitude system (Dehaene 1997; Hauser 2000). Thus values can be compared and combined approximately, but there is a lot of room for slop (Weber's Law error), as we will see.

A second source of complexity in values is that there turn out to be many sorts of value, each of which pertains to different entities and plays its own role in rules of inference. My impression is that cross-disciplinary discussions of value have often foundered because psychology deals primarily with one sort, economics with another, and moral philosophy with yet another, and ordinary language conflates them, calling them all "value."

A third source of complexity is that most of these sorts of value appear in two different versions, which I will call the *objective* and the *subjective* versions. In the objective version, the judger attributes value to something in his or her conceptualized world: *X is of value*. In the subjective version, the judger attributes value to something in his or her conceptualized world, relative to some observer (who may be the judger him- or herself): *X is of value to Y/to me*. This distinction appeared already in the discussion of psychological predicates in chapter 7, and we will build on it here.

A final difference between values and points is that the rules of inference internal to the value system encompass many more possibilities than simple addition. In working these rules out, I hope to establish that they are intuitively plausible, since after all they are meant to capture basic generalizations about how we reason with values.

In this chapter, we will work through the different kinds of value and some of the basic inferences that they license. Just to see where we are headed, here is a brief description of them:

- *Affective value (A-value)* A situation or action has A-value for an individual X if it yields pleasure or suffering, if it *feels* good or *feels* bad to X.
- *Utility (U-value)* A situation or action has U-value for an individual X if it yields a benefit or exacts a cost, if it is good *for* X or bad *for* X. I will often lump A-value and U-value together as *A/U-value*.
- *Resource value (R-value)* An object has R-value if it is *valuable*, if it is good *for someone to have*.
- *Quality (Q-value)* The Q-value of an object or action is measured relative to other objects or actions of the same type, usually in terms of its function. Thus we speak of a good/excellent computer or a bad/poor back dive.
- *Prowess (P-value)* The P-value of a person is measured by quality of performance at some task: the person is good/excellent *at* doing such-and-such or bad/poor *at* doing such-and-such.
- *Normative value (N-value)* N-value concerns conformity to social norms, including moral/ethical norms, religious norms, and cultural norms such as customs, manners, and etiquette. A person's action has N-value to the extent that it conforms to norms. We say it was good/right *of* X to do such-and-such or bad/wrong *of* X to do such-and-such.
- *Personal normative value (PN-value)* A person has PN-value to the extent that he or she conforms to social norms. We speak of a person with positive PN-value as being *good* or *virtuous*, and of one with negative PN-value as being *bad* or *wicked.*
- *Esteem (E-value)* E-value concerns the overall social value of a person and represents a composite of PN-value, P-value, dominance, and other factors. We speak of a person with high E-value as being *prestigious* or *respected.*

We will take these up in turn.

9.2 Affective Value

Affective value is the most visceral kind of value, the one most rooted in biology. A situation or action has A-value to an individual if it yields pleasure or suffering, if it feels good or bad. Stereotypical A-good situations include eating a tasty meal, enjoying an aesthetic experience, being in pleasurable company, having good sex, and so on. Stereotypical A-bad situations include being hungry, sick, in pain, tied up, among enemies, in danger, and so on. Questions of preference, likes and dislikes, and approach and avoidance can be couched in terms of the A-value of the situations and actions involved. Thus this is the kind of value of greatest concern to certain strains in psychology (e.g. Herrnstein 1993; Mandler 1993).

One's judgments of the A-value of a situation to oneself are linked to the character of one's own experience. As in the previous three chapters, this observation returns us to the valuation features of consciousness, discussed in section 3.3. Recall that the perceptual systems give the contents of consciousness their *form*, while the valuation features give them their *feel*—the sense of familiarity or novelty, the sense of reality versus imagination, the sense of being self- versus world-controlled, and so on. The valuation feature relevant to A-value is *[affective]*, which comes with a positive or negative valence. If the cognitive structure of an entity contains the feature [+affective], this entity is experienced as something that *matters*, either positively or negatively, depending on the valence. Thus a situation that comes with the valuation [+affective: valence±] can be judged to have a positive or negative A-value.

Note that I am making a distinction between *experiencing* a situation as attractive or aversive (i.e. having an experience that involves the feature [±affective]) and *judging* its A-value. There are two reasons. First, one can make judgments of A-value that distance one from experience: "That may look like fun, but I know better—it's really painful." Such judgments require *reasoning* about A-value, which has to be done in terms of the conceptual system, not in terms of valuation features per se. Second, the affective valuation feature at best connects only to one's own experience. It cannot account for the ability to attribute likes and dislikes to others, and to compare one's own with theirs: "That meal tasted good to her, but not to me"; "It was good for me; was it good for you?" That is, judging the A-value of a situation to someone else goes cognitively beyond one's own experience; it clearly involves theory of mind.[2]

A formalization of A-value appears in (1). Notice that the constraint on the second argument, *Animate/Person*, is identical to the constraint on holders of psychological predicates (chapter 7) and situational and actional attitudes (chapter 8). This follows from the fact that A-VAL too is a basic psychological predicate.

[2] However, one may experience a proxy for someone else's feeling through empathy (which of course may or may not be veridical). Empathy is a different kind of theory of mind, a less conceptual one than what is under discussion here.

(1) *A-value of a situation*
A-VAL ([$_{\text{Situation}}$ S], [$_{\text{Animate/Person}}$ X]) = valence × magnitude
'The A-value of situation S to animate/person X is valence times magnitude.'
(where valence ranges over + and −; I will often omit the magnitude)[3]

This leads to analyses like (2).

(2) a. Fred enjoys eating figs.
A-VAL ([FRED EAT FIGS], FRED) = +
'The A-value to Fred of eating figs is positive.'
b. Joe dislikes being beaten.
A-VAL ([SOMEONE BEAT JOE], JOE) = −
'The A-value to Joe of someone beating him is negative.'

Notice how A-VAL more or less replicates the valence associated with psychological predicates in section 7.7. There we treated these predicates as simplex (e.g. BORED, INTERESTED, HAPPY) and just appended the valence as an intuitive add-on (BORED is −, HAPPY is +, and so on). A-VAL presents the possibility of making the treatment of valence more systematic, by making A-VAL a component of all these predicates; but I will not do it here.

As with the psychological predicates, with A-values there is a form in which the Experiencer is syntactically suppressed, and the value is presented as though objective: *Eating figs is enjoyable* parallels the "objective" *That problem is interesting*. The same analysis offers itself: in these examples, the Experiencer is the implicit generic individual notated by YA. (Alternatively, a particular Experiencer may be suggested by the context—for example 'you' in *Was that enjoyable?*)

(3) a. Eating figs is enjoyable.
A-VAL ([YA EAT FIGS], YA) = +
b. Being beaten sucks.
A-VAL ([SOMEONE BEAT YA], YA) = −

[3] A more traditional notation might use a function A-VAL′ of three variables that yields a truth-value, along the lines of (i).

(i) [A-VAL′ (SITUATION, PERSON, valence × magnitude)]

The function A-VAL in (1) can be derived from this by lambda-abstraction on the last variable. However, since we eventually want to be able to compare and add values, the form in (1) is more convenient.

The reasons for having both subjective and objective forms are the same as with the psychological predicates (section 7.8), and the ways one shifts between them in reasoning are the same as well. In the interests of space, I will not repeat the arguments here.

9.3 Utility

Turning to utility, or U-value: an event or situation has U-value for someone if it yields a benefit or exacts a cost. A simple expression of U-value is *good/bad for so-and-so*, as in (4).

(4) a. Eating this broccoli will be good for Bill.
    b. Being overweight is bad for Max.

The same event may be of positive U-value and negative A-value; for instance, eating broccoli may be good for Bill, but Bill may hate eating broccoli. Likewise, the same event may simultaneously be beneficial to one person and harmful to another; for example, the action of *revealing the name of the thief* might be good for the police and bad for the thief. Hence utility is a function not of the event per se but of the event's effect on its participants (or even on bystanders).

As with A-value, there are expressions of utility that do not name a participant, for instance (5a,b), in which the assertion of value presents itself as an objective property of the event or situation.

(5) a. Eating broccoli is good.
    b. Being overweight is bad.

And again the understood participant can be an implicit generic individual, as in *Eating broccoli is good for ya*.

The formalization of U-value is exactly like that for A-value.

(6) *Subjective U-value*
    a. Eating broccoli is good for Bill.
       U-VAL ([BILL EAT BROCCOLI], BILL) = +
    b. Being overweight is bad for Max.
       U-VAL ([MAX BE OVERWEIGHT], MAX) = –

    Objective U-value
    c. Eating broccoli is good.
       U-VAL ([YA EAT BROCCOLI], YA) = +
    d. Being overweight is bad.
       U-VAL ([YA BE OVERWEIGHT], YA) = –

Now comes a crucial point. Following the overall mentalist tenets of the present approach, *objective A-value and U-value are still not value in the world, independent of observers.* You may think that it's objectively good to eat broccoli, and I may not. But each of us conceptualizes this value as a property of eating broccoli. What makes the value objective in the present sense is that we disagree about the value of eating broccoli independent of any particular person. By contrast, if you think eating broccoli is good for Bill, and I think it's not good for Harry, we have no disagreement: we're talking in terms of subjective U-value.

Because A-value and U-value function pretty much the same in the larger system of values, I will often lump them together as A/U-value.

9.4 Getting Into and Out Of the System

For a system of values to be of use, there have to be principles for assigning values to situations—the "input rules" mentioned in section 9.1. There also have to be principles that use the values of one's potential actions in order to decide what to do—"output rules." Just using A-value and U-value, we can begin to see the general outlines of these principles.

An important input principle comes from the macrorole tier of chapters 6 and 7. One macrorole function we discussed was X AFF$^{\pm}$ Y, 'X affects Y positively/negatively'. From this we can draw an inference: being a Patient, who is negatively affected, either is a negative experience or exacts a cost; being a Beneficiary, who is positively affected, either is a positive experience or yields a benefit. The principle can be stated as (7).

(7) *Tuning of A/U-value to valence of AFF*

$$\begin{bmatrix} [\textit{any thematic tier}] \\ \text{(X) AFF}^{\alpha}\ \text{Y} \end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{A/U-VAL}\ (\beta, \text{Y}) = \alpha$$

(where α ranges over $+$ and $-$)

Let's unpack this. The left-hand side of the rule is an Event or Situation, with the macrorole function X AFF$^{\pm}$ Y. Hence the character Y is a Patient if the valence is negative, and a Beneficiary if the valence is positive. The right-hand side is a value judgment over a situation β, which is in turn bound to (thus identical with) the situation on the left-hand side of the rule. The valence of AFF is bound to the valence of the value judgment: if the valence is positive, the situation is of positive A/U-value to Y; and if the valence is negative, the situation is of negative value to Y. In short, it's nice to be helped and yucky to be victimized.

The other macrorole function is X $\text{EXP}^{\pm}$ Y, 'X has a positive/negative experience of Y'. This too gives us an inference to a value—this time to A-value only, as we are speaking of the character of experience.

(8) *Tuning of A-value to valence of EXP*

$$\begin{bmatrix} [\textit{any thematic tier}] \\ \text{X EXP}^{\alpha}\ (\text{Y}) \end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{A-VAL}\ (\beta, \text{X}) = \alpha$$

In parallel to (7), (8) says that 'an Event or Situation that X experiences positively is of positive A-value to X; an Event or Situation that X experiences negatively is of negative A-value to X'. The upshot of these two rules, then, is that the character of an event in which a person is involved can lead to a judgment of that event's A/U-value to that person.

There have to be other sources of U-value besides (7). For one thing, taking almost *any* action entails some sort of cost, but only some actions return a tangible benefit. For another case, let's go back to the example of eating broccoli. It perhaps is judged to have negative A-value because eating broccoli entails *tasting* broccoli, an unpleasant experience (for some people). But the judgment of its U-value must be derived from knowledge of nutrition or the like.

These input principles are the counterpart of the rules of a game that say "Doing such-and-such gives you so many points." We next need output principles that are the counterpart of "You win the game if you have the most points." About the simplest possibility is that the A/U-value of a potential action affects its preferability; that is, people prefer to do actions that are better for them.

Here is a version of the principle. I have to state it informally, because many of its pieces have not yet been formalized (but the treatment of actional attitudes in chapter 8 gets us close).

(9) If X is considering two actions, Act_1 and Act_2, and X wishes to
commit to (come to intend) one of them, and
A/U-VAL ([X ACT_1], X) > A/U-VAL ([X ACT_2], X), then
X will come to intend Act_1.

This is a prediction about people's behavior, including one's own.

There also has to be a procedural version of the rule, by which one directs one's own actions, perhaps (10).

(10) *(Procedural version)*
If EGO is considering two actions, Act_1 and Act_2, and wishes to
commit to (come to intend) one of them, and

A/U-VAL ([EGO ACT$_1$], EGO) >
A/U-VAL ([EGO ACT$_2$], EGO), then
INTEND Act$_1$

The special font in (10) is meant to indicate a procedural instruction: as it were, Act$_1$ is put in the queue for action. One isn't thinking about it any more, one is preparing to do it. Perhaps (10) can be interpreted as an interface principle that connects conceptualization to the action system sketched in chapter 4.

(9) and (10) differ in an interesting way. (9) represents a primitive version of a "folk theory of rational choice"; it involves theory of mind, because we are making judgments about what other people judge valuable and what they intend. (10), on the other hand, is a primitive account of *how we actually choose*, based on A- and U-value. It does not involve theory of mind, since the only mind it applies to is one's own.

In section 4.4, we touched upon the difficulties in formulating a more robust version of (10), when we asked how one chooses a repair strategy if an action such as making coffee breaks down. The solution involved comparing the relative cost (U-value) of the alternatives, but this was acknowledged as formidable. In particular, it takes for granted the dangerous fiction that A/U-values are comparable—that they can be placed on a linear scale. This is the place where the present approach connects with rational choice theory, but also with issues like Nisbett and Wilson's (1977) illusions of choice, Kahneman, Slovic, and Tversky's (1982) heuristic principles of choice, Gigerenzer et al.'s (2000) "fast and frugal strategies," and Tetlock's (2003) "taboo trade-offs." This is also the place where the theory incorporates the time-dependence of values. For instance, one case of "giving in to temptation" is choosing an immediate gratification (an action with positive A-value) in preference to a potentially greater A-value or U-value to be realized over a longer time-span. (See Ainslie 2001 and Stevens and Hauser 2004 for discussion of this issue of "temporal discounting.")

In addition to A- and U-values, normative values also go into the mix. One may choose to "do the right thing" (i.e. the action with higher N-value) despite its higher cost (U-value) or unpleasantness (A-value). Alexander (1987) says that what we mean by a "saint" is someone who always makes such choices, altogether disregarding A-value. Most of us are not saints, and thus we often yield to temptation in choosing an action with high A-value but negative N-value.

Doris and Stich (2005) observe that (in present terms) knowing the N-values of actions is not the same as committing to these N-values. They cite evidence that some psychopaths fail to make this extra step: they know what the rules are but do not feel bound to observe them. Committing to N-values involves taking them into account in calculating what to do, that is, incorporating them into rule (10).

The main point is that (9) and (10), or suitably amplified versions of them, are principles that serve as "output rules" to the value system. (9) predicts what people will do based on their judgments of value (or it can be used in reverse, to retrodict their values from what they do). And (10) directly tells you what to do.

9.5 Resource Value, Quality, and Prowess

9.5.1 Resource Value

A third type of value is what I will call *resource value* (or R-value). An object has R-value if it is *good for someone to have*; a simpler expression is just that the object is *valuable*. One reason something may be good to have is that it offers the potential (or affordance) for an event with A- or U-value. For a simple case, food has R-value because it offers the potential of being eaten, which in turn is an action of A- and U-value to the eater. Similarly, a house has R-value because it offers the potential of being lived in. Another reason something may be good to have is that it adds to the esteem (E-value) of its owner, as in the R-value of a famous painting or a fancy car or fashionable clothes (according to whatever happens to define "fashionable" in the wearer's social milieu).

For another prominent case, money has R-value to its holder because it offers the potential of being exchanged either for other objects with R-value or for the performance of actions with A/U-value to the holder. Economists, whose basic data are exchanges, are therefore most concerned with R-values and those A/U-valued actions for which exchanges for R-valued objects are possible, namely labor and services (e.g. Akerlof and Yellen 1993; Scitovsky 1993). A full discussion of the sources of R-value is well beyond the scope of this book. However, sections 10.1 and 10.5, which deal with fairness in distribution of resources and with exchanges, offer entries into the issues.

R-value, like A/U-value, comes in subjective and objective varieties. All the examples so far are objective, in the sense that the object simply has R-value rather than R-value *to so-and-so*. The subjective/objective contrast is hard to express using the word *good*, but it turns up in expressions like (11a,b).

(11) a. This piece of land is very valuable/worth a lot to Harry. (subjective)
b. This piece of land is very valuable/worth a lot. (objective)

(11a) leaves the question open of whether the land means anything to anyone else. By contrast, in (11b) the sense is that anyone will value the land highly; that is, the valuer is the implicit generic individual YA, as usual.

The formalization of R-value is straightforward.

(12) a. *Subjective R-value*
R-VAL ($[_{\text{Object}}$ Y], $[_{\text{Person}}$ X]) = valence × magnitude
'The R-value of object Y to person X is such-and-such.'
b. *Objective R-value*
R-VAL ($[_{\text{Object}}$ Y], YA) = valence × magnitude
'The R-value of object Y (to anyone) is such-and-such.'

Normally, the valence of an R-value is positive. However, it may be useful to say that a debt has a negative R-value.

R-value is defined in terms of its affordance for A/U-valued actions. But there is a secondary interaction between R-value and A-value. To the extent that having available resources reduces anxiety, the situation of having things with R-value can itself be of A-value: having stuff feels good; lacking stuff feels bad. Of course, the strength of this interaction varies from person to person.

9.5.2 Quality

An object or event can be valued in terms of its *quality* (Q-value) relative to other objects or events of the same type.

(13) a. This is a good/terrible computer. (object)
b. That was an excellent/miserable back dive. (action)

Typically, when objects are rated for quality, it is in terms of a particular action for which the object is to be used.

(14) a. This spatula is good for frosting cakes with.
b. This book is good for learning about deconstructionist phonetics.

When the *for*-phrase is absent (*a good spatula*, *a good book*), there is still an implicit purpose, as observed by Katz (1966) and Pustejovsky (1995). The default interpretation is that the object in question has quality with respect to performing its *proper function*—what the object is *for* (in the sense of Millikan 1984, or the concept's "telic quale" in the sense of Pustejovsky 1995). A good spatula is one that is good for scraping and spreading viscous materials (usually food-related), and a good book is one that is good for reading.

The adjective *excellent* is natural in expressions of Q-value; by contrast, it is somewhat awkward in expressions of A-value.

(15) a. This spatula is excellent for frosting cakes with.
b. ??Drinking milk is excellent for you.

An extension of Q-value concerns the use of some object for the function normally played by something else (Aronoff 1980). Expressions like (16a), in particular the *makes a good X* construction, are characteristic of this reading.

(16) a. This rock is/makes a good table.
b. This table is/*makes a good table.

Alternatively, the purpose may be inferred from conversational context: for instance, in the context of comparing clouds for their resemblance to cows, one might say *THAT cloud's good.*

Further cases of Q-value might arise with adjectives that describe attractiveness, such as *beautiful* (+) versus *ugly* (−), and with a pair like *strong/weak*, which describes ability to apply or resist force. Millikan (1984) in a sense proposes to think of the truth-value of a sentence as its Q-value: true is positive Q-value, false is negative Q-value.

As I won't be making use of Q-value in any of the formal rules of this chapter, I won't formalize it here.

9.5.3 Prowess

Related to Q-value, there is a type of value that rates the quality of an individual's performance, as in (17a). A second syntactic form, (17b), parallels the attribution of Q-value to artifacts, as in (13a) and (17c). We might call the sense in (17a,b) *prowess* or P-value. Note that like quality, prowess can be expressed by using the adjective *excellent*.

(17) a. Harry is good/excellent (at singing). (prowess)
b. Harry is a good/excellent singer. (prowess)
c. This is a good/excellent knife. (quality)

Like the Q-value of an artifact, P-value is in effect an affordance for some task. One demonstrates P-value by performing actions of Q-value.

9.6 Normative Value and Personal Normative Value

The most complex type of value is *normative value* (or N-value), which concerns conformity to social norms of the sorts discussed in chapter 5. Among the subvarieties of N-value are moral/ethical value, religious value, and value according to standards of custom, fashion, and etiquette (manners and politeness). Among these, moral/ethical value is the main sort of value of interest to philosophers (e.g. Stich 1993; Harman 2000). Unlike the previous sorts of value, N-value is strongly situated in the social domain: it has to do not just with people, but with people in the context of social interaction.

As observed in chapter 5, the subvarieties of N-value share a great deal of their linguistic expression and often apply in similar ways to similar situations, but they can still be teased apart into separate subdomains. For example, it is possible for a highly moral person to have bad manners; conversely, a person with exemplary manners may well be deeply immoral. And either of these combinations may be combined with religiosity or its absence. For the purposes of this chapter, they all behave about the same, so for the most part I will lump them all together. If desired, however, the different subvarieties of N-value can be notated with a subscript on N, so that moral value is notated N_{moral}-*VAL*, manners $N_{manners}$-*VAL*, and so on. To some extent, these have characteristic predicates: *virtuous* and *evil* go with moral N-value, *polite* and *ill-mannered* with the N-value for manners, *pious* with religious N-value, and so on. *Right* and *wrong* go with several subvarieties.

On the other hand, I adamantly want to avoid conflating N-value with other sorts of value, in a way often seen in folklore and even the philosophical literature. The story of Cinderella presents a stereotypical case. The virtuous heroine (high N-value) is also beautiful (high Q-value), and the wicked stepsisters (low N-value) are also ugly (low Q-value), as if this correlation were the most natural thing in the world. Similarly (if I may be impertinent toward one of the greats), Plato often identifies the beautiful (Q-value) with the good (N-value). Moreover, he constantly asks how to define a good man, intending personal normative value, the consequence of performing N-valued actions. But then he pursues the answer by asking what makes someone a good doctor or a good pilot or a good harpist, all issues of P-value.[4] I take such conflation to be a mistake.

Characteristic expressions of normative value judgments are shown in (18). (18a,b) foreground the action; (18c) foregrounds the actor.

(18) a. It is good of Harry to wash the dishes without being asked.
b. Washing the dishes without being asked is good of Harry.
c. Harry is good to wash the dishes without being asked.

As with A/U-value, the attribution of N-value is focused on a relation between a person and an event. However, this time the event *must* be something that the person does intentionally, as in (19a); it cannot be something that happens to the person, as in (19b,c), or that the person does accidentally, as in (19d). By contrast, U-value can be ascribed to such situations (20).[5]

(19) *N-value*
a. Washing the dishes is good of Harry.
b. *Being overweight isn't good of Harry.
c. *Being appointed chairman was good of Harry.
(except if this means he was good to *allow* himself to be
appointed chairman, which is an intentional action on his part)
d. *Getting lost on the way home wasn't good of Harry.

[4] Plato justifies this approach by correctly saying that a good doctor is good at healing people, and so forth, which is precisely the definition of P-value in terms of proper function. He makes the case of a good man comparable by simply asserting (in present terms) that the proper function of a human being per se is to be virtuous. But this entirely begs the question. It's not clear to me that a human being per se is conceptualized as having a proper function any more than a rabbit or a tree does. (I am grateful to Nancy Bauer for bringing this issue to my attention.)

[5] *Should* and *ought to* exhibit ambiguity parallel to that of *good.* The predictive sense (i) does not have to do with values. But the prudential sense (ii) expresses utility, and the normative sense (iii) expresses normative value. The reader can verify that the normative sense has the same constraints on intention as normative *good.*

(i) The bus should/ought to arrive soon. (predictive)

(ii) You should/ought to take an umbrella in case it rains. (prudential: 'The U-value of taking an umbrella is positive.'; 'It would be good for you to take an umbrella.')

(iii) You should/ought to wash the dishes. (normative: 'The N-value of washing the dishes is positive.'; 'Washing the dishes would be good of you.')

(20) *U-value*
a. Being overweight isn't good for Harry.
b. Being appointed chairman was good for Harry.
c. Getting lost on the way home wasn't good for Harry.

We can formalize the representation of N-value in the usual way. (21a) is the general expression, and (21b) gives an example.

(21) *Subjective N-value*
a. N-VAL ($[_{\text{Situation}, +\text{Action}}$ X ACT], $[_{\text{Person}}$ X]) = valence × magnitude
'The N-value of X's doing such-and-such is valence times magnitude.'
b. Washing the dishes is good of Harry.
N-VAL ([HARRY WASH DISHES], HARRY) = +

As with A- and U-value, it is possible to omit the person from an attribution of N-value, as in (22). As usual, if some contextually relevant Actor is not intended, this implies that it would be good or bad of *anyone* to perform the action, and so the formalization incorporates the generic YA as the actor. This can be thought of as the "objective" version of N-value.

(22) *Objective N-value*
a. It is good to wash the dishes without being asked.
Washing the dishes is good.
N-VAL ([YA WASH DISHES], YA) = +
b. It is bad to kill people.
Killing people is bad.
N-VAL ([YA KILL PEOPLE], YA) = –

There is also another way to bleach out the relational character of N-value: it is possible to omit the action from an expression of N-value, as in (23a). The sense is then that the person's generic actions, whatever they may be, are of N-value. This manipulation is impossible with A/U-value: a paraphrase like (23b) in terms of A/U-value makes no sense.[6]

[6] Note the parallel between the four ways of expressing N-value/PN-value and the four ways of expressing psychological predicates pointed out in chapter 7. This further points up their conceptual similarity.

(i) (= (19b))
Washing the dishes is good of Harry. This story bores Harry.

(ii) (= (23a))
Washing the dishes is good. This story is boring.

(23) Harry is good.
  a. = ‘Harry does things of positive N-value.’
  b. ≠ ‘Things happen of positive U-value for Harry.’

I will formalize this as a further kind of value, *personal normative value* (PN-value).

(24) Dick is evil.
  PN-VAL (DICK) = –

9.7 Some Inferences Involving Normative Value

Next we have to add input rules to this part of the value system, that is, principles that assign N-values to actions. Every culture has a huge collection of such principles: “It’s normatively good to dress in such-and-such a way,” “It’s normatively good to shake hands (or bow, or whatever) upon greeting someone with whom one is in a formal relationship,” “It’s normatively bad to eat such-and-such,” “It’s normatively bad to use such-and-such words,” “It’s normatively bad to show attraction to people of the same sex,” and so on. These all have to be learned in all their detail either through teaching or through observation. As noted in chapter 5, learning these is a major component of learning a culture.

What appears to be a much more general input rule arises from an interaction of A/U-value with N-value. For example, actions on your part such as charity, which are good *for* someone else (A/U-value to them), are considered good *of* you (N-value). Conversely, gratuitous violence is bad *of* the perpetrator (N-value) and bad *for* the victim (A/U-value). The general principle is that it’s N-good to do things that affect others positively and N-bad to do things that affect them negatively. (This is the principle that has been investigated most extensively in the recent experimental tradition seeking moral universals (e.g. Hauser 2006).)

(25) $$\begin{bmatrix} \text{X ACT} \\ \text{A/U-VAL } (\beta, \text{Y}) = \alpha \end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{N-VAL } (\beta, \text{X}) = \alpha$$
‘If X acts in a way that is good/bad *for* Y, then that action is good/bad *of* X.’

(iii) (= (19c))
Harry is good to wash the dishes. Harry is bored with this story.

(iv) (= (24a))
Harry is good. (PN-value) Harry is bored.

Let me take a little care in decoding the notation here, as it will recur profusely throughout the rest of this chapter and chapter 10. The upper line on the left-hand side of (25) identifies an action by X. The lower line is a modifier that says that the A/U-value of some action β to Y is α. In turn, β is bound to X's action. Thus the left-hand side encodes an action by X whose A/U-value to Y is α.

The right-hand side of (25) says that action β, the action performed by X, has an N-value. This N-value is bound to the A/U-value on the left-hand side. Thus if X's action has a positive effect on Y, then it is good of X to do it; if X's action has a negative effect on Y, then it is bad of X to do it. Thus (25) opens up a vast range of entries into the N-value system: anything anyone does that affects anyone else is potentially a target for assigning an N-value.

The next question is, how do N-values affect one's action? That is, what are the output rules for N-value? Put differently, why should one do (N-)good things and avoid doing (N-)bad things?

One reason appeared in section 9.4: N-value contributes to calculating the preferability of an action in comparison to other possible actions (rules (9)–(10)). In other words, (9)–(10) serve as output rules for N-value as well as for A- and U-value.

Another interaction is that it may *feel* good/bad (A-value) to *do an act* that's good/bad (N-value): that is, performing an act with N-value may result in an accompanying secondary A-value, which in turn contributes to its preferability. In common language, we say that someone who experiences this interaction "has a conscience." In particular, when the value is negative, I think the secondary A-value is called the feeling of *guilt*. This is a principle internal to the value system that moves from N-value to A-value. In turn, the A-value feeds into the output rules (9)–(10) for deciding what to do.

Notice that the very same event may thereby receive values from multiple sources. For example, illicit sex may be of positive A-value, deriving from the inherent character of the experience itself. At the same time, it receives negative N-value, and (for some people, anyway) the secondary negative A-value of guilt—that's why we call it "illicit." The balance of these factors in deciding whether or not to act varies from individual to individual, and for each individual, from context to context.

So far the effect of N-value on behavior has been through the A/U-value system. But there is an additional and less direct possibility. The basic intuition is that if you do something good, it makes you a good person; if you do something bad, you're a bad person. That is, performing actions with N-value affects your PN-value. This is a rather peculiar principle, but, strikingly, it conforms to intuition. A first version is (26).

(26) $\text{N-VAL (X ACT, X)} = \alpha \Rightarrow_{\text{default}} \text{PN-VAL (X)} = \alpha$
'If it's N-good/bad of X to do some action, then X is PN-good/bad.'

We might call (26) the "absolute" verson of the principle.

A slightly more sophisticated or "relative" version of this principle might say that good acts add to your total "goodness" and bad acts subtract from it. Such a rule cannot be stated within a static logic, which presumes a timeless database. Rather, as with rule (10), it is necessary to introduce a dynamic or procedural logic, one that allows values to be updated as one acts over time. Such a system will be necessary in any event in order to allow for belief revision (i.e. deciding one was wrong about something). Without being very specific about how such a system works, we might state the rule in question as something like (27), where ADD TO is meant as a procedural instruction.

(27) $\begin{bmatrix} \text{X ACT} \\ \text{N-VAL } (\beta, \text{X}) = \alpha \end{bmatrix}^{\beta} \Rightarrow \textsf{ADD } f \times \alpha \textsf{ TO } \text{PN-VAL (X)}$
(where α = valence × magnitude, and f is a multiplier on the magnitude)
'If X acts in a way that's good/bad of X, that makes X a better/ worse person.'

The consequence of (27) is that a person's PN-value at any moment is related to his or her history of performing normatively valued actions.

The result of the "addition" in (27) is not necessarily a sort of mathematical sum, since values are measured by the analogue magnitude system, which is far less rigid than arithmetic. Furthermore, the relative weights of current and past actions have a highly subjective character, which is reflected in the "wild card" multiplier f. We intuitively recognize this subjectivity in statements like these:

(28) a. I've now performed so-and-so many good actions. Is that enough to make up for all the bad things I've done?
b. Even though what you just did was horrible, I'm not going to hold it against you.
c. That one horrible thing you just did has wiped out my whole good opinion of you.

Exactly what determines the effect of a current action on PN-value under different circumstances is a question far beyond the scope of the present exploration. ((28c) has the flavor of contamination, along the lines of Nemeroff and Rozin's (2000) proposals about food contamination: a speck of bad stuff, say a cockroach leg, contaminates a whole lot of food.)

Notice how PN-value parallels prowess (P-value). Just as one acquires high P-value by performing high-quality actions (high Q-value), so one acquires high PN-value by performing highly N-valued actions.

Just another reminder: rules (26)–(27) are not meant to treat X's PN-value as a freestanding thing unto itself. We are not describing how X acquires "real" value; rather, we are describing how the judger conceptualizes X acquiring "objective" value. Lest this should seem a problematic stance on PN-value, it should be recalled that this is the very same stance taken in studying vision, where perception is described in terms of the perceiver developing a conceptualization of the objective "world out there," in response to certain inputs to the perceptual system. In both cases, we are concerned with the individual's sense of what is real. And from this point of view, goodness is as real as size.

Doris and Stich (2005) distinguish two principal approaches to ethics: identifying morally obligatory actions, and identifying "what sort of person to be," so-called virtue ethics. In the present framework, these approaches divide clearly into a focus on N-value versus a focus on PN-value. They point out that people tend to draw an inference that (PN-)good people do (N-)good things, and (PN-)bad people do (N-)bad things; that is, (26) is taken to be a two-way entailment. This leads to the possibility of entirely opposite rationalizations of people's actions, depending on one's opinion of them: "The president is bad, so whatever he does, no matter how harmless it looks, must have a pernicious goal behind it" versus "The president is good, so whatever he does, no matter how superficially bad it looks, must be in our best interests." (The latter case is also used in reasoning about God, of course.) Doris and Stich observe that this folk bias tends to be carried over into virtue ethics: a person can be considered virtuous only if he or she can do no wrong. Yet, as they point out, almost everyone is corruptible given the wrong situation (and they cite experimental evidence to this effect), and even "morally weak" people sometimes do the right thing for the right reasons. This suggests that virtue ethics suffers from taking PN-value to be a strictly binary distinction rather than, like all other measures of value, a graded and possibly multidimensional one.

(26)–(27) draw inferences from N-value to PN-value, but they are still internal to the value system as a whole, and they do not yet affect one's course of action. Next we have to ask, why does it matter to be a good person? To answer this question, we have to step into our final domain of value, esteem.

9.8 Esteem

Esteem (E-value), like PN-value, pertains specifically to persons. Subjective E-value is expressed by verbs such as *respect* and *esteem*; objective E-value is expressed by predicates like *prestigious*, *esteemed*, and *reputation*. Since having a good reputation implies that *anyone* will respect you, the implicit generic YA is again the valuer in objective E-value.

(29) a. *Subjective E-value*
Joe respects Harry.
E-VAL (HARRY, JOE) = +
b. *Objective E-value*
Harry is prestigious/well-respected.
Harry has a good reputation.
E-VAL (HARRY, YA) = +

Esteem seems to be based on a composite of personal normative value, prowess, status in the dominance hierarchy, wealth (accumulation of R-value), personal attractiveness (Q-value), and perhaps other factors as well. We can state this intuition roughly as (30).

(30) $\text{E-VAL (X, Y)} = c_1 \times \text{PN-VAL (X)} + c_2 \times \text{P-VAL (X)} + \ldots \text{other factors}$
(where c_i is a normalizing constant, possibly context-dependent)
'X's esteem depends on X's virtue, X's prowess, and other factors.'

The ratio of importance among the factors contributing to E-value (represented in (30) by the normalizing constants) may be highly variable and context-dependent. Moreover, since we are dealing with the analogue magnitude system in measuring values, the idea of a normalizing constant in the usual mathematical sense is far too specific. In any event, one kind of value can outweigh others. For example, to be *notorious* is to have a high E-value, owing to prowess or wealth, despite a negative PN-value (think of Al Capone).

One of the "other factors" in (30) is group membership: by default, one accords members of one's own group higher esteem than members of other groups, and one also accords greater esteem to members of higher-status groups (whatever their other qualities). However, in spite of this default, particular individuals of low-status groups who have other highly respected qualities (virtue, prowess, leadership, etc.) may be accorded greater respect than the default (think of Gandhi).

A related factor in (30) is "You're known by the company you keep": people adjust an individual's esteem in terms of the esteem of the people he or she associates with. "Oh, my! YOU know Noam Chomsky?" (This can be positive or negative, depending what you think of Chomsky.) This factor has a bit of the flavor of contagion, along the lines discussed by Nemeroff and Rozin (2000), though unlike contagion it's not necessarily biased toward the negative. Thus another way one can affect one's E-value is through social climbing.

(30) begins to show a reason why it might matter to be a good person (i.e. to be of high PN-value): it raises the esteem in which one is held. Still, the same old question comes up: what difference does that make to one's life? That is, what are the output rules that convert E-value into behavior?

The interesting thing is that the main effect of E-value is on the behavior of others. The most obvious of such behavior is *displaying* respect or disrespect. There are many ways of doing this, and they vary drastically from culture to culture. Some of the more prominent and universal ways of displaying respect are giving gifts, extending hospitality, deferring to the other's choice, expressing a compliment, and making culture-specific gestures such as shaking hands and bowing. Some of the more prominent and universal ways of displaying disrespect are declining to share, ignoring the other's preferences, and expressing an insult, either verbally or through culture-specific gestures (often with sexual connotations) like Bronx cheers and the finger (Eibl-Eibesfeldt 1989). The whole huge component of cultural knowledge dealing with manners and etiquette appears to be aimed at governing displays of respect and disrespect, and it may infuse every action involving another individual.

Many of these behaviors have primate antecedents, particularly in the expression of dominance. This leads to the observation that E-value need not be confined to human societies. Of course, some components of it, such as PN-value and wealth, are purely human. (Well, maybe there's even a chimpanzee counterpart of wealth: consider the temporary respect shown to someone who happens to have a lot of meat to share.) But other components such as dominance and prowess certainly play a role in chimpanzee societies and might be considered to constitute a sort of proto-E-value. Then the human version just adds further components to a preexisting conceptual system.

Let me formalize the demonstration of respect as (31): it is an action by X that makes X's respect for Y observable.

(31) X displays (dis)respect for Y.
$\text{X DISPLAY [E-VAL (Y, X)} = \alpha]$

In the terms of section 5.5, (31) is the structure on the social plane that is linked to the physical action of displaying on the physical (observable) plane: it is what makes the physical action count as a symbol of respect or disrespect. To the extent that nonhumans can be shown to have the concept and use it to modulate their behavior (e.g. choosing when to display), we can say that a version of (31) is a component of their dominance and submission displays.

Now it is an important wired-in part of our minds that it feels good to be shown respect, and bad to be shown disrespect; and these feelings reveal themselves in gestures and postures that, as pointed out by Darwin, have nonhuman antecedents. This makes clear the nature of the macrorole tier associated with (31): in displaying respect, X affects Y positively (Y is a Beneficiary); in displaying disrespect, X affects Y negatively (Y is a Patient). So this is another one of those rules where valences are co-tuned.

(32) Y displays (dis)respect for X.
$$\begin{bmatrix} \text{Y DISPLAY [E-VAL (X, Y)} = \alpha] \\ \text{Y AFF}^{\alpha}\text{ X} \end{bmatrix}$$

In turn, by rule (7), Y's display is of A-value to X, which is why it matters to X—which in turn is why it matters to X to be a good person.

A second effect of E-value on how people treat you is discussed by Alexander (1987), Fehr and Fischbacher (2004), and many others: an individual of high esteem is sought out by others for cooperative interactions of mutual benefit. (33) is an approximation; ACT WITH might be replaced with the formal treatment of joint tasks in section 8.8.

(33) $\text{E-VAL (X, Y)} = \alpha \Rightarrow_{\text{default}} \text{U-VAL ((Y ACT WITH X), Y)} = \alpha$
'If Y holds X in high/low esteem, it's good/bad for Y to associate with X.'

Again, to the extent that X's being held in esteem leads to opportunities for cooperation, and cooperation brings X benefits, esteem is a U-good thing to have. That is, esteem is a sort of personal R-value.

The upshot is that unlike A- and U-value, which directly affect one's own action, E-value is Machiavellian, in that it is basically a way of manipulating the actions of others. One effects this manipulation by performing actions that enhance one's E-value in the eyes of others, in particular N-valued actions. Thus normatively valued actions connect to the rest of the system in at least three ways: it may feel good/bad to do them, it may be "the right/wrong thing" to do them, and doing them may increase or decrease one's E-value. Alexander (1987) observes that, as a consequence of this last factor, people are always trying to get everyone else to judge them altruistic (N-good) while concealing their actual lack of altruism.

9.9 Summary

Table 9.1 sums up the types of value discussed here.

Here are the principles involved in using the value system that have been mentioned in the course of the chapter, including the example numbers of those principles that have been formalized.

Input rules

- Some situations and actions feel good or bad (have an intrinsic A-value).
- Actions have a cost and/or benefit (U-value).
- Having something good or bad happen to one (including an act on the part of someone else) has an A/U-value. (7)–(8)
- Huge numbers of actions are assigned N-value as part of cultural knowledge.
- One's E-value takes into account one's group membership and one's dominance.

Output rules

- People will do what has the best combination of A-value, U-value, and N-value for them. (9)
- DO what has the best combination of A-value, U-value, and N-value. (10)

Rules internal to the system

- It may feel good (A-value) to have things of R-value, and it may feel bad to lack things of R-value.
- It may feel good (A-value) to do something of positive N-value, and it may feel bad (A-value) to do something of negative N-value.
- It is of positive N-value to do something of positive A/U-value for someone else, and it is of negative N-value to do something of negative A/U-value to someone else. (25)

Table 9.1
Varieties of value (entity to which value is ascribed is underlined)

Type of value	Applies to ontological type	Subjective version	Objective version
Affective (A-)value	Events, situations	Situation $\underline{X}$ feels good to Y	Situation $\underline{X}$ feels good
Utility (U-value)	Events, situations	Situation $\underline{X}$ is good for Y	Situation $\underline{X}$ is good
Resource (R-)value	Objects	Object $\underline{X}$ is valuable to Y	Object $\underline{X}$ is valuable
Quality (Q-value)	Events		Event $\underline{X}$ was a good one
	Objects		Object $\underline{Y}$ is good for doing X
Prowess (P-value)	Persons		$\underline{Y}$ is good at doing X $\underline{Y}$ is a good X-er
Normative (N-)value	Action + person	Doing $\underline{X}$ is good of $\underline{Y}$	Doing $\underline{X}$ is good
		It is good of $\underline{Y}$ to do $\underline{X}$	It is good to do $\underline{X}$
		$\underline{Y}$ is good to do $\underline{X}$	
Personal normative (PN-)value	Persons		$\underline{X}$ is good/virtuous
Esteem (E-value)	Persons	X respects $\underline{Y}$	$\underline{Y}$ is prestigious

- Doing something of positive/negative N-value affects one's PN-value positively/negatively. (26)–(27)
- One's E-value takes into account one's PN-value, one's P-value, perhaps one's personal attractiveness (Q-value), and the E-values of those one associates with. (30)
- It is of positive/negative N_{manners}-value to display positive/negative respect (E-value) for someone else.
- Displaying (dis)respect (E-value) for someone has A-value for that person. (32)+(7)
- There is U-value in cooperating with individuals of high E-value. (33)

This is all beginning to feel like it has the requisite degree of complexity and richness for a framework in which to describe cultures in some detail.

Overall, then, the value system can be seen as an abstract calculating system that helps govern action. The system assigns values to actions or anticipated actions, and uses these values eventually to determine the personal value of individuals performing these actions. In turn, the personal value of these individuals is crucial in determining how to interact with them. The value system is multidimensional, in that there are at least eight kinds of value, most with objective and subjective versions; and each type plays a different role in the system of inferences built on value. My impression is that previous approaches have been limited because they insist on a unitary notion of value, and because in many cases (especially in moral philosophy), they discount subjective value and culture-specific value altogether (see discussion in section 5.10).

Crucial to the present approach is that values are being treated as part of a cognitive system: every individual capable of social interaction is competent with this system. What is necessary in learning a culture is acquiring the rules that assign values to particular sorts of action: what it's good to do and what it's bad to do, which actions fall under moral prescription and which under (mere) manners, how and when one is to display respect, and so on.

In the context of how other cognitive systems are now understood—especially language—this analysis should not be too surprising. The complexity of the system is not undermined by the fact that value judgments are often quick and intuitive. In language, judgments of grammaticality and meaningfulness are quick and intuitive: the computational reasons for these judgments are deeply unconscious. In vision, judgments of spatial configuration and motion are intuitive and present themselves to awareness as "what is the case in the world." Thus value judgments are of a piece with the rest of cognition. However, unlike what we find with linguistic and visual judgments, aspects of value judgments are available to awareness as well—they lie on the borderline between intuitive and conscious reasoning.

This chapter has also begun to connect the value system to its linguistic expression. Part of the job of linguistic semantics is to explicate the meanings of words and phrases. In the domain of spatial language, linguistic semantics has benefited from the attempt to develop formal analyses of the conceptualization of space, motion, force, and agency in terms of a limited conceptual vocabulary and combinatorial system, and large portions of language have been subsumed under such analyses. The present exploration has undertaken a similar analysis for the sizable parts of the vocabulary whose meanings incorporate notions of value.

Finally, this approach offers the possibility of making many longstanding questions more precise. At the scale of the individual lifespan: what is the course of development of value systems in humans (Piaget 1932; Kohlberg 1981–84; Turiel 1983; Macnamara 1991; Premack and Premack 1994; Bloom 2004)? Over historical time: to what extent are value systems a functional outcome of what it takes to make a society work well (and interact well with other societies) (Fiske 1991; Jacobs 1994)? Over evolutionary time: how much of the basis of human value systems is innate? And of that, how much is part of our primate heritage, and how much is unique to humans (de Waal 1996; Hauser 2000, 2006)? By taking a formal overview of the entire system, it is possible to pose these questions in a more comprehensive context.

I don't want to pretend that the analysis here is an ultimate solution. For the moment, it is worth recognizing how much descriptive breadth and depth has been achieved in a relatively short chapter, and how many potential connections to other disciplines have been drawn, while at the same time acknowledging that this is only a first step.

Chapter 10

Fairness, Reciprocity, and Exchange

This chapter uses the notions of value developed in chapter 9 to work out formulations of the basic social notions of fairness, reciprocity, and exchange. In particular, I will draw a strong distinction between freely chosen reciprocity and agreed-upon exchange, two concepts that often seem to be conflated in the literature on reciprocity. I will also bring out a variant of reciprocation in which one reciprocates by displaying esteem, rarely mentioned in any but the anthropological literature.

The discussion is fairly heavy on the use of formalism; I will do my best to give useful glosses to help keep track of what is going on.

10.1 Fairness and Selfishness

Intuitively, an action is *fair* if it is equally good or bad for everyone. As can be seen by the use of *good* ***for*** and *bad* ***for***, the sense of good and bad implied by this intuition is utility (U-value): the action provides the same benefit to everyone or exacts the same cost from everyone. The action might also involve A-value: everyone gets the same pleasure or suffering from the action. Thus we can formalize this intuition as (1).

(1) Y acts fairly toward $X_1, \ldots, X_n$.
For all X_i, X_j:
A/U-VAL ((Y ACT), X_i) = A/U-VAL ((Y ACT), X_j)
'Y's act is as good for X_i as it is for X_j.'

Such a situation can be played out in various ways. Y's act may be a single action that impinges on everyone at once. Or it may be a composite of multiple subactions at different times, each impinging on a different individual: if you do such-and-such to X_1 this time, you'd better do something of equal value to X_2 the next time.

There are two ways to determine whether an action is fair, depending on how one calculates its value to the participants. I will call these *subjective* and *objective* construals of fairness. The subjective construal is to find out how each X individually values Y's action and make sure these are equal. But this is labor-intensive, even if one can compare subjective values. In particular, it may involve lots of use of theory of mind, always a strain. There is a shortcut, though. Recalling the discussion of subjective versus objective evaluation in section 7.8: if you don't know the value of something to somebody, the default is to assume it's the objective value—that is, the value to a generic individual. In terms of this shortcut, all individuals count as the same, and the action is fair if everyone gets the same thing. This is the objective construal of fairness.

To see how these versions of fairness compare, consider for example an action where Y is distributing resources to the group. In the simpler objective construal, each person in the group receives the same amount of resources. In the more sensitive subjective construal, the differing needs of individuals are taken into account. Similarly, when the action is collecting resources from the group, the simpler construal is a uniform tax, and an approximation of the more sensitive construal is a progressive tax. (Which counts as "fairer"? It depends how you frame the issue, as we know from the political domain.)

The form of (1) may make it look as though Y is an individual outside the set of Xs. But this need not be the case. Y may in fact be one of the Xs, in which case Y is *sharing* resources equally with the other members of the group. (However, I find it strange to apply the term "fair" to a situation in which Y is a member of a group, and Y's action is something that hurts all the Xs as much as him- or herself.) Another possibility is that Y is the entire set of Xs, and the action is a joint task for the benefit of the group—preparing a communal meal, raising money for a playground, or the like. In this case, what counts as fair allocation is equal cost and equal benefit. In the case of playing a game, what is at stake is equal opportunity for participants, perhaps guaranteed by taking turns. On the objective construal, everyone does the same and everyone gets the same benefit; the subjective construal permits more modulation.

Two related patterns of behavior are acting selfishly and acting altruistically, which can be analyzed like this:

(2) a. Y acts selfishly toward $X_1, \ldots, X_n$.
For all X_i:
A/U-VAL ((Y ACT), Y) > A/U-VAL ((Y ACT), X_i)
'Y's action is better for him- or herself than it is for anyone else.'

b. Y acts altruistically toward $X_1, \dots, X_n$.
For all X_i:
$\text{A/U-VAL}((\text{Y ACT}), X_i) > \text{A/U-VAL}((\text{Y ACT}), \text{Y})$
'Y's action is better for everyone else than it is for him or her.'

Another prevalent pattern is distribution by rank: an action fits this pattern if the higher your rank, the better the treatment you get. Let's use *OUTRANK* to stand for an inequality of personal value—either prowess (P-value), dominance, virtue (PN-value), or general esteem (E-value). Then the principle can be stated as (3).

(3) Y acts according to distribution by rank among $X_1, \dots, X_n$.
For all X_i, X_j:
X_i OUTRANKS $X_j \Rightarrow$
$\text{A/U-VAL}((\text{Y ACT}), X_i) > \text{A/U-VAL}((\text{Y ACT}), X_j)$
'If X_i outranks X_j, then Y's act is better for X_i than it is for X_j.'

This mode of distribution is appropriate for awarding honors and prizes. But it's of far broader social application: the top dog receives the best place at the table, the best mate, the most food, the plum work assignment, and (if he or she transgresses) the most lenient punishment, while those on the lowest rungs get the fewest resources, the most unpleasant work, the most severe punishment, and so on. This principle is of course amply attested in social animals, where OUTRANK is realized in terms of the dominance hierarchy.

With these pieces in hand, we turn to the central thesis of Alan Fiske's *Structures of Social Life* (1991): the hypothesis that human societies have exactly four ways to distribute goods, labor, and responsibility, and that these four ways are universal innate structures within the human social cognitive capacity. Cultures differ not in having one of these structures rather than another, but in how they distribute the use of the four structures over different contexts. Fiske arrives at this hypothesis through painstaking analysis of a vast number of social institutions in disparate cultures.

Fiske's four "elementary forms of human relations" are the following:

- In *Communal Sharing*, each member of a group shares equally in benefits and responsibilities, or, within limits, "From each according to his abilities, to each according to his needs." The prototypical case is distribution of food at a family meal (in our culture, at any rate).
- In *Authority Ranking*, benefits and responsibilities are distributed according to rank.

- In *Equality Matching*, equality among group members is guaranteed by each member doing or receiving exactly the same thing. Prototype cases are turn-taking and voting.
- In *Market Pricing*, participants exchange resources and/or labor according to negotiated agreement.

I would be inclined to admit a fifth elementary form, which Fiske does not recognize:

- In *Competition*, each participant is trying to obtain more benefits and avoid more costs than the other.[1]

The basic pieces of these frames are evident in what we have already discussed: distribution of benefits and costs across a group. Let's try to formalize them.[2] Each of the frames can be considered a norm: "One should distribute benefits, costs, and responsibilities in such-and-such a way." That is, each frame assigns a positive N-value to a particular sort of distribution. Since the four norms often conflict with each other, one's culture has to play an important role by contextualizing the frames: "In the following sorts of activities [insert culturally specific list here], one should distribute benefits, costs, and responsibilities in such-and-such a way." Thus yet another important part of learning a culture is learning the list of activities that goes with each frame.

Viewed this way, Communal Sharing might be encoded as (4). It is a norm that places positive value on acting fairly in the sense of (1). It doesn't say *how* to act fairly; rather, it gives fairness a positive normative value. To sidestep the notational complexity of quantifying over groups of arbitrary size, I will state (4) and subsequent principles in terms of groups containing only two members. (I leave a fully quantified version as an exercise for masochistic readers.) The category ACT_{CS} is the class of actions that the culture defines as subject to Communal Sharing. The membership of this category must be learned.

[1] Competition is different from Fiske's "asocial" relation, in which participants simply ignore each other's existence (the stance we often take toward people on the bus), in that each participant is tracking the other's actions for opportunities to take advantage. Section 5.8 suggested that competition is the "evil twin" of cooperation.

[2] Fiske himself formalizes the differences among the frames in terms of different and incommensurable mathematical systems, an approach that I find questionable in cognitive terms. The present treatment strikes me as closer to the right approach, because it is grounded in independently necessary notions of value and joint action.

(4) *Communal Sharing (CS)*
For actions in the category ACT_{CS} and a group $\{\text{X}_1, \text{X}_2\}$:
$$\text{N-VAL}\ (\begin{bmatrix}\text{YA ACT}_{\text{CS}} \\ \text{A/U-VAL}\ (\alpha, \text{X}_1) = \text{A/U-VAL}\ (\alpha, \text{X}_2)\end{bmatrix}^{\alpha}, \text{YA}) = +$$
'It's N-good to act in a way that is equally A/U-good for X_1 and X_2.'

Principle (4) can be taken as a version of Rawls's (1971) doctrine of "justice as fairness," and it undergirds the notion of "equality under the law." Notice however that actual moral and legal systems, in adopting a principle of equality, often circumscribe who counts as an X: only members of one's own group, only men, only white men, only white men who own property, and so on. Furthermore, actual cultures always delimit the category of actions to which Communal Sharing applies. A culture is considered more egalitarian to the extent that this category of actions and the category of Xs is broader.

Consider again the case in which the actor is the group as a whole, doing a joint task. Here (4) creates normative pressure for each individual to pitch in equally and not shirk, so that the costs in labor or resources are spread evenly. Of course, it's up to each individual how much to yield to this pressure. Moreover, the "sensitive" construal of the rule may license one to say, "Well, so-and-so has more time to give than I do" or "Well, so-and-so cares more about this cause than I do," justifying giving less effort without feeling that one has violated the norm.[3]

Competition is of course the flip side of Communal Sharing: the goal is to grab as much for oneself as possible and make matters as difficult for the other as possible. (5) states what it means to be selfish in a dyadic interaction; it is a reduced form of (2), the overall statement of what it means to be selfish.

(5) X acts selfishly toward Y.
$\text{A/U-VAL}\ ((\text{X ACT}), \text{X}) > \text{A/U-VAL}\ ((\text{X ACT}), \text{Y})$
'X's action is better for him- or herself than it is for Y.'

A normative stance of selfishness or competitiveness is easily derived from (5); the normative value in (5) is allowed to be either positive or

[3] Fiske proposes other important manifestations of Communal Sharing, such as collective ethnic identity, sense of group unity, and "losing one's identity in the group." I am inclined to see these as symptoms of group membership per se (see section 5.7), rather than aspects of Communal Sharing. They are connected with Communal Sharing in the sense discussed here because the typical domain for Communal Sharing is the social group.

neutral (i.e. "It's OK to..."). ACT_{COMP} is the class of actions that the culture deems appropriate for competition. Y_{COMP} is the class of individuals against whom it is culturally appropriate to compete; this class is likely to include any member of another group, since the normal mode of interaction with other groups is competition.

(6) *Competition (COMP)*

For actions in the category $\mathrm{ACT_{COMP}}$ and individuals $\mathrm{Y_{COMP}}$:

$$\text{N-VAL}\ (\begin{bmatrix}\text{YA ACT}_{\text{COMP}} \\ \text{A/U-VAL}\ (\alpha, \text{YA}) > \text{A/U-VAL}\ (\alpha, \text{Y}_{\text{COMP}})\end{bmatrix}^{\alpha}, \text{YA}) = +/0$$

'It's N-good/N-OK to act in a way that is better for yourself than for Y.'

Next consider Authority Ranking, which is a norm that places a positive value on rank-based distribution. This can be formalized as (7), where the category ACT_{AR} consists of those actions for which the culture condones Authority Ranking.

(7) *Authority Ranking (AR)*

For actions in the category $\mathrm{ACT_{AR}}$ and a group $\{X, Y\}$:

$$\text{N-VAL}\ (\begin{bmatrix}\text{YA ACT}_{\text{AR}} \\ \text{X OUTRANKS Y} \Rightarrow \\ \quad \text{A/U-VAL}\ (\alpha, \text{X}) > \text{A/U-VAL}\ (\alpha, \text{Y})\end{bmatrix}^{\alpha}, \text{YA}) = +$$

'It's good to act in a way that reflects individuals' relative status/merit.'

Authority Ranking is what justifies bosses getting paid more than workers, rewarding or honoring individuals for merit, giving first prizes that are bigger than second prizes, and so on. What's important here is that, because Authority Ranking is the operative norm in these situations, such disparities are acceptable even to low-ranked individuals.

On the other hand, Authority Ranking does not sanction dominant individuals taking resources from subordinates by force; that is, it is not synonymous with oppression. Oppression instead falls under the negative side of rule (26) of chapter 9, "It's N-bad to be mean to people." I repeat the rule here:

(8) $$\begin{bmatrix}\text{ACT(X)} \\ \text{A/U-VAL}\ (\beta, \text{Y}) = \alpha\end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{N-VAL}\ (\beta, \text{X}) = \alpha$$

'If X acts in a way that is good/bad *for* Y, then that action is good/ bad *of* X.'

It is possible that much of the delicate dynamic between dominants and subordinates is a consequence of the interplay of (7) and (8). Although entitled to more resources and respect, a dominant does not want to be construed as oppressive, which would result in lower PN- and E-value. He or she is therefore motivated to show a bit more respect and generosity to subordinates, rather than insisting too heavily on Authority Ranking. (Another normative principle that can limit the freedom of dominants is "You shouldn't think too highly of yourself," a principle that emerges at the end of section 10.3.)

Let us next turn to Equality Matching. The present context enables us to see a distinction between Equality Matching and the previous two norms, not (to my knowledge) noticed by Fiske: all the cases he cites, such as voting to elect an official, taking turns helping to harvest a field or carpool the children, and participating in a rotating credit association, are joint tasks. Thus, in order to formalize Equality Matching, let us first review the treatment of joint tasks from section 8.8. (9) repeats (68) from chapter 8 (using INTEND instead of COM for mnemonic clarity).

(9) X and Y intentionally perform action A jointly.

$$\begin{bmatrix} \{X, Y\}\ \mathrm{ACT_A} \\ \text{COMPOSED-OF } \{[X\ \mathrm{ACT_x}], [Y\ \mathrm{ACT_y}]\} \\ [\text{FROM } [\{X, Y\}\ \text{INTEND } \alpha]] \end{bmatrix}^{\alpha}$$

The first line says that X and Y are joint actors in action A. The second line says that X's part in this is $\mathrm{ACT_x}$ and Y's part is $\mathrm{ACT_y}$. The third line says that the joint task arises from X and Y's jointly intending to perform the task α, where α is bound to the joint action as a whole.

Given this formalization, Equality Matching can be stated as (10). ACT_{EM} is the class of actions that the culture considers appropriate for Equality Matching.

(10) *Equality Matching (EM)*

For actions in the category $\mathrm{ACT_{EM}}$ and a group {X, Y}:

$$\text{N-VAL}\left(\begin{bmatrix} \{X, Y\}\ \mathrm{ACT_{EM}} \\ \text{COMPOSED-OF } \{[X\ \mathrm{ACT_x}], [Y\ \mathrm{ACT_y}]\} \\ [\text{FROM } [\{X, Y\}\ \text{INTEND } \alpha]] \\ X\ \mathrm{ACT_x} = Y\ \mathrm{ACT_y} \\ \text{A/U-VAL}(\alpha, X) = \text{A/U-VAL}(\alpha, Y) \end{bmatrix}^{\alpha}, \{X, Y\}\right) = +$$

'It's good for everyone in a joint task to do exactly the same thing and benefit equally.'

The large bracketed expression here requires decoding. The first three lines are the definition of a joint task, copied from (9). The fourth line says that X and Y participate in the action in the same way. The fifth line says that X and Y benefit equally from the joint action. (10) as a whole says that a joint action that satisfies all these conditions is normatively good of the participants.

Equality Matching might be considered a special case of Communal Sharing, in which (a) the actor is the group as a whole, and (b) the value to the participants is measured strictly by the objective construal, so there is not only equal effect on the members of the group, but also equal participation. In this respect, it is the most rigid of Fiske's four frames—yet its rigidity makes it an excellent way to coordinate certain sorts of joint actions. In the usual cases where it involves turn-taking, in effect it is institutionalized reciprocation.

Fiske's fourth frame, Market Pricing, pertains to the dynamics of exchange transactions, so I will postpone its treatment until after we discuss exchanges in section 10.5.

10.2 Reciprocation, Retaliation, and Restitution

Consider the relation between two actions that is expressed by a certain use of the preposition *for* in English.

(11) a. Fred cooked Lois dinner *for* fixing his computer.
b. Fred slashed Lois's tires *for* insulting his sister.

These sentences describe situations in which someone does something *in return* for someone else's action. (11a) describes an action with a positive value; (11b) describes an action with negative value. Such acts of reciprocity can felicitously take place only with another person, an entity that can be regarded as having values and responsibility. One cannot sanely punish one's computer for crashing.

If we switch around the actions between the examples in (11), we get sentences that sound odd or perhaps ironic.

(12) a. #Fred cooked Lois dinner for insulting his sister.
b. #Fred slashed Lois's tires for fixing his computer.

This shows that we expect a positively valued action in return for a positively valued action, and a negatively valued action in return for a negatively valued one. Reciprocation is further sensitive to the (analogue) magnitude of values as well: we find it odd if the two actions related by *for* do not match in magnitude. The sentences in (13) convey some of this oddness.

(13) a. #Fred cooked Lois dinner for saying hello to him.
(overreaction)

b. #Fred cooked Lois dinner for rescuing all his relatives from certain death. (underreaction)

c. #Fred slashed Lois's tires for eating too little at dinner. (overreaction)

d. #Fred slashed Lois's tires for murdering his entire family. (underreaction)

In (13a,c), we sense Fred as overreacting, as doing something unwarranted in return for Lois's action; in (13b,d), we sense him as underreacting, as doing something that is not nearly enough to recognize the importance of Lois's action.

The intuition, then, is that a reciprocal action calls for rough equivalence of value between the two actions. Crucially, a particular action may be of different value to the participants, and the original actor need not even know that his or her action helped or harmed the reciprocator. Thus the principle of reciprocation must be stated in terms of the particular person the action affects, that is, subjective A/U-value.

In order to formalize the conditions on reciprocation, first we must introduce a modifier on an action, called *RECIP*, which can be glossed as 'in return for'.

(14)
$$\begin{bmatrix} \text{Y ACT}_2 \\ \text{RECIP [X ACT}_1] \end{bmatrix}$$

'Y performs ACT_2 in return for X performing ACT_1.'

RECIP has a strict inference on the temporal relation between the two actions. This is shown in (15), where T_1 and T_2 are the times at which the two actions take place.

(15)
$$\begin{bmatrix} \text{Y ACT}_2\text{; T}_2 \\ \text{RECIP [X ACT}_1\text{; T}_1] \end{bmatrix} \Rightarrow T_1 < T_2$$

'A reciprocal action takes place after the action that it reciprocates.'

We can now state the principle that lies behind the judgments of appropriateness in (11)–(13). Like many inferences with value, this one is defeasible.

(16) *Principle of reciprocation*

$$\begin{bmatrix} \text{Y ACT}_2 \\ \text{RECIP [X ACT}_1]^{\alpha} \end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{A/U-VAL}(\beta, \text{X}) = \text{A/U-VAL}(\alpha, \text{Y})$$

'When Y acts in return for X's acting, Y's act is as good/bad for X as X's act is for Y.'

Decoding this: the upper line on the left-hand side says that Y performs some action, ACT_2. The modifier on the bottom line says that this action is in reciprocation for X's performing a different action, ACT_1. The right-hand side says that the value of β (= Y's action) to X equals the value of α (= X's action) to Y.

The logic of reciprocity expressed by (16) encompasses a behavioral strategy much discussed in the ethological literature, *reciprocal altruism*. This is sometimes phrased as "You scratch my back and I'll scratch yours" (e.g. Dawkins 1989). However, this terminology actually suggests not reciprocal altruism, but an explicit agreement to perform a joint action in the sense of section 8.8. The two scenarios have different inferences: it is *nice* to reciprocate altruistically, but there is no necessity of doing so, and there is no necessary communication between the participants. By contrast, a joint action calls for verbal or nonverbal agreement, and once agreement is reached, one is *obligated* to perform one's role. Thus a better phrasing for reciprocal altruism is "I'll scratch your back *because* you scratched mine." Although various cases of reciprocal altruism are documented in nonhumans (and these are not without question; see Stevens and Hauser 2004), humans are distinguished by the generality with which they can use all sorts of actions in reciprocation.[4]

However, (16) is also broader than reciprocal altruism, because it leaves open whether the value in question is positive or negative. If the value is negative, (16) expresses the principle of retaliation (or retribution), which is certainly found in nonhumans to some extent, for example in the vervet monkeys studied by Cheney and Seyfarth (1990). In the human case, the equivalence of values amounts to a more or less formal statement of "The punishment fits the crime": this helps guide what responses are appropriate in retaliation for harmful actions.

Principle (16) tells you that if you do something bad to someone, you may expect retaliation. However, you may be able to head retaliation off

[4] I think it's important to keep the terminology clearer than it typically is here. For instance, reciprocal altruism is often studied formally in terms of Iterated Prisoner's Dilemma (e.g. in Dawkins 1989). However, in Iterated Prisoner's Dilemma there is by assumption no communication between the participants; they can only observe and respond to the others' moves in the game. Thus it is misleading to speak of them as ever "cooperating": cooperation requires a joint intention. Real cooperation in Prisoner's Dilemma would be a situation in which the two participants get to make an explicit deal to maximize their collective benefits. Under such a construal, "defecting" goes beyond just acting selfishly: it consists of welshing on the deal—breaking one's obligation to the other player.

by performing another kind of reciprocation for negative actions, illustrated in (17). Notice that the *for* of reciprocation is used again, but in this case it can be filled out as *to make up for* instead of *in return for*.

(17) a. Fred cooked Lois dinner (to make up) for having embarrassed her in public.
b. Fred brought Lois flowers (to make up) for forgetting her birthday.

Here the perpetrator of the negative action is performing a positive action in *restitution*, righting the balance. One might consider *reconciliation* in primates (Cheney and Seyfarth 1990) to be an evolutionary ancestor of restitution. Except in the legal literature, I find little discussion of restitution and its relation to reciprocation.

As with reciprocation and retaliation, restitution requires a rough equivalence of value: notice the weirdness of (18).

(18) a. #Fred gave Lois his vast fortune (to make up) for forgetting her birthday. (overreaction)
b. #Fred brought Lois flowers (to make up) for killing her whole family. (underreaction)

Unlike reciprocation, restitution is not neutral to the valence of the original action: there is no counterpart that reverses the signs, as might be expressed by a sentence such as (19).

(19) ##Fred slashed Lois's tires (to make up) for remembering her birthday.

Thus the principle of restitution can be stated as (20). The bottom line on the left-hand side encodes the restriction illustrated in (19): it stipulates that X's original action must have negative value for Y.

(20) *Principle of restitution*

$$\begin{bmatrix} \text{X ACT}_2 & \\ \text{REST} & \begin{bmatrix} \text{X ACT}_1 \\ \text{A/U-VAL}\ (\alpha, \text{Y}) = - \end{bmatrix}^{\alpha} \end{bmatrix}^{\beta} \Rightarrow_{\text{default}}$$

$$\text{A/U-VAL}\ (\beta, \text{Y}) = -\text{A/U-VAL}\ (\alpha, \text{Y})$$

'When X acts in restitution for having done something bad for Y, the restitutive act is as good for Y as the original act was bad for Y.'

Notice that if X does something harmful to Y, one possible sort of retribution for Y is to force X to perform restitution, perhaps through the intervention of the authority of the group. However, forced restitution is not always equivalent to retaliation. For one extreme case, "Nothing you can make the murderer do will bring my son back"; that is, restitution is not possible, even though retaliation might be (e.g. killing the murderer's son).[5]

(16) and (20) are stated as inferences from events of reciprocation and restitution to the value of the actions in question. However, this is not enough: one *should* reciprocate actions that benefit one and one *should* perform restitution for having harmed others; that is, there is a normative value attached to these actions. How this value plays out—what actions should be reciprocated and restituted, and what counts as appropriate reciprocation and restitution—is another variable among cultures and subcultures. But the overall principle seems universal. (21a,b) state these principles. They are special cases of the positive version of rule (8): "It's good to be nice to people."

(21) a. *Normative value of reciprocal altruism*

$$\begin{bmatrix}\text{X ACT}_1\\ \text{A/U-VAL }(\beta,\text{Y}) = +\end{bmatrix}^{\beta} \Rightarrow \text{N-VAL }(\begin{bmatrix}\text{Y ACT}_2\\ \text{RECIP }(\beta)\end{bmatrix},\text{Y}) = +$$

'If X does something that is good for Y, it's good of Y to reciprocate.'

b. *Normative value of restitution*

$$\begin{bmatrix}\text{X ACT}_1\\ \text{A/U-VAL }(\beta,\text{Y}) = -\end{bmatrix}^{\beta} \Rightarrow \text{N-VAL }(\begin{bmatrix}\text{X ACT}_2\\ \text{REST }(\beta)\end{bmatrix},\text{X}) = +$$

'If X does something that is bad for Y, it's good of X to perform restitution.'

The left-hand side specifies that X's original action is of value to Y: positive in (21a), negative in (21b). The right-hand side says that reciprocation by Y or restitution by X is normatively good.

What about retaliation? Depending on the circumstance, there are three possible normative principles. One is "You should retaliate," generalizing the left-hand side of (21a) to negative valence, as in (22a). A second is "It's all right to retaliate," which differs from (22a) only in that the normative value, instead of being positive, is greater than or equal to zero. The third is "You should turn the other cheek," in which case the right-hand side places a negative N-value on retaliation, as in (22c).

[5] The distinction between these two in conceptual development is noted by Piaget (1932); he claims that the notion of restitution is established later than that of retribution.

(22) a. *Normative value of retaliation is positive*

$$\begin{bmatrix}\text{X ACT}_1 \\ \text{A/U-VAL } (\beta, \text{Y}) = - \end{bmatrix}^{\beta} \Rightarrow \text{N-VAL } (\begin{bmatrix}\text{Y ACT}_2 \\ \text{RECIP } (\beta)\end{bmatrix}, \text{Y}) = +$$

‘If X acts in a way that harms Y, it’s good/right of Y to retaliate.’

b. *Normative value of retaliation is neutral*

$$\begin{bmatrix}\text{X ACT}_1 \\ \text{A/U-VAL } (\beta, \text{Y}) = - \end{bmatrix}^{\beta} \Rightarrow \text{N-VAL } (\begin{bmatrix}\text{Y ACT}_2 \\ \text{RECIP } (\beta)\end{bmatrix}, \text{Y}) \geq 0$$

‘If X acts in a way that harms Y, it’s OK for Y to retaliate.’

c. *Normative value of retaliation is negative (“Turn the other cheek”)*

$$\begin{bmatrix}\text{X ACT}_1 \\ \text{A/U-VAL } (\beta, \text{Y}) = - \end{bmatrix}^{\beta} \Rightarrow \text{N-VAL } (\begin{bmatrix}\text{Y ACT}_2 \\ \text{RECIP } (\beta)\end{bmatrix}, \text{Y}) = -$$

‘If X acts in a way that harms Y, it’s bad/wrong of X to retaliate.’

Not only do these three principles conflict with each other, but in addition (22a,b) conflict with rule (8) (“It’s N-bad to hurt someone”), while (22c) does not. Again, a lot of cultural variation arises from how these rules are understood to apply in practice. I suspect that it’s partly in the service of negotiating such conflicts that more explicit moral and legal codes arise.

There are a number of ways for slippage to be introduced into reciprocation. Perhaps the most pernicious arises from a general cognitive bias toward overestimating harm to oneself and underestimating harm to others. Hence, if I retaliate against you, you judge the harm done to you to be greater than the harm you originally did to me. You are therefore motivated to even the score by retaliating further, leading to escalating cycles of violence.

10.3 Honoring, Shaming, and Apologizing

Another circumstance in which English uses the *for* of reciprocation has received little discussion in the literature on reciprocal altruism. However, it bears a striking parallel to the sorts of reciprocation discussed in the previous section. This use appears in the examples in (23).

(23) a. Joe praised Sue *for* saving the drowning child.

b. The club honored Sue *for* her service to the community.

c. The chairman awarded Sue a gold medal *for* winning the race.

d. The fans cheered Sue *for* hitting a grand slam in the ninth inning.

In these cases, Sue may or may not have done anything to benefit the individual or organization that is acting reciprocally. Rather, she has done something of high N-value (23a,b) or Q-value (23c,d). These increase her

PN-value and prowess (P-value), respectively. In turn, by rule (30) of chapter 9, PN-value and P-value both contribute to the esteem in which Sue is to be held.

In this light, we can see all the actions in (23) as displays of respect: they are performative actions such as honoring and thanking, which have the effect of making esteem publicly observable. As discussed in section 9.8, such displays of respect are naturally of positive A-value to Sue as well; and the resource value of increased esteem is a further benefit. The function of the *for* of reciprocation in (23) is to express that these displays are undertaken in response to particular esteem-raising actions on Sue's part.

The negative counterpart of (23) is shaming: humiliation and community-sanctioned punishment for normative transgressions and low-quality actions—even when these actions do not directly harm anyone else. These are displays of *dis*respect. (24) gives two examples.

(24) a. Sue scolded Bill for his bad manners. (N-value transgression)
b. The fans booed Foulke for giving up a grand slam in the ninth inning. (Q-value transgression)

If (23) parallels reciprocal altruism, and (24) parallels retaliation, we expect a parallel also of restitution. And there is one: *apology* is restitution for harm done, by displaying self-humiliation before the injured party.

(25) Foulke apologized to the manager for giving up a grand slam.

(26) formalizes these sorts of reciprocation, using the notion DISPLAY proposed in section 9.8.

(26) a. *Normative value of honoring someone who does something estimable*

$$\begin{bmatrix}\text{X ACT}\\ \text{N/Q-VAL }(\beta) = +\end{bmatrix}^{\beta} \Rightarrow$$

$$\text{N-VAL }(\begin{bmatrix}\text{Y DISPLAY (E-VAL (X, Y)} = +)\\ \text{RECIP }(\beta)\end{bmatrix}, \text{Y}) = +$$

'If X does something honorable or estimable, it's good of Y to express positive esteem for X for having done so.'

b. *Normative value of apologizing to someone you've hurt*

$$\begin{bmatrix}\text{X ACT}\\ \text{A/U-VAL }(\beta, \text{Y}) = -\end{bmatrix}^{\beta} \Rightarrow$$

$$\text{N-VAL }(\begin{bmatrix}\text{X DISPLAY (E-VAL (X, X)} = -)\\ \text{REST }(\beta)\end{bmatrix}, \text{X}) = +$$

'If X does something bad to Y, it's good for X to express negative self-esteem to Y to make up for having done so.'

As with retaliation, there are conflicting possibilities for shaming/humiliation, played out differently in different circumstances in different societies (and even among different individuals). The formalizations of these possibilities are clear by analogy with (22) and (26a), and I will not state them here.

In these cases of reciprocal displays of respect, it is hard to know what counts as equivalence in value between the original act and the reciprocal act. Perhaps the best one can do is context-dependent proportionality (parallel to authority ranking): first prize ought to be more valuable than second prize; greater praise should be accorded to someone who saves 80 lives than to someone who stops on the highway to help you fix your car; a bigger faux pas calls for more fervent contrition.

A curious point: notice that etiquette (a type of N-value) demands that the recipient of one of the reciprocal actions in (26) respond "I don't deserve it; what I did was nothing." Why is this the case? The immediate cause seems to be a principle to the effect that you shouldn't think too highly of yourself. (27) gives two possible versions of this principle.

(27) a. N-VAL ([E-VAL (X, X) >> 0], X) = –
'It's bad of X to esteem him- or herself much above zero.'; 'You shouldn't rate your esteem very high.'; 'Don't have too high an opinion of yourself.'

b. N-VAL ([E-VAL (X, X) > E-VAL (X, YA)], X) = –
'It's bad of X to esteem him- or herself above his or her objective esteem.'; 'You shouldn't rate your esteem higher than it really is.'

I leave it an open question whether this principle can be derived from more basic principles. In any event, it can serve as a pressure toward egalitarianism.

Again, the fact that these phenomena involving displays of respect so closely parallel the previous cases of reciprocation is an unexpected result that emerges from examining linguistic data; it thus serves as an interesting vindication of the methodology adopted here.

10.4 Deserving

Now we come to some very peculiar but pervasive reasoning about values. In the interests of clarity and compactness, I will go through this part of the argument without formalization.

Let's look at a number of ways to state the normative principle for reciprocation. (28a) is closest to the form in rule (21a); (28b) is a different phrasing. In both cases, the normative principle applies to the person Y who has benefited from X's original action.

(28) *Reciprocation*
If X does something A/U-good for Y,
a. then it is N-good of Y to reciprocate. (= (21a))
b. then Y should reciprocate/reward X.

However, the situation can also be construed in a stronger sense: not only would it be good for Y to reciprocate, Y has a sort of "moral obligation" to reciprocate, expressed perhaps as (28c). If we turn around and look at this situation from X's point of view, we might express it as (28d,e).

(28) *Reciprocation, continued*
c. then Y owes it to X to reciprocate.
d. then X should be rewarded by Y.
e. then X deserves to be rewarded by Y. ("One good turn deserves another.")

A counterpart involving restitution appears in (29).

(29) *Restitution*
If X does something A/U-bad to Y,
a. then it is N-good of X to provide restitution. (= (21b))
b. then X should provide restitution.
c. then X owes it to Y to provide restitution.
d. then Y should be compensated by X.
e. then Y deserves compensation from X.

If the normative system condones retaliation, then the counterpart involving retaliation is (30).

(30) *Retaliation*
If X does something A/U-bad to Y,
a. then it is N-good/N-OK of Y to retaliate. (= (22a))
b. then Y should/may retaliate.
c. then Y is entitled to retaliate.
d. then X should be punished by Y.
e. then X deserves to be punished by Y.

These inferences (or whatever they are) are altogether common in our reasoning.

So far so good. But then these serve as stepping-stones to further normative conclusions we typically draw, whether or not they are logically warranted. Suppose we put ourselves in the place of the individual who deserves something in the (d) and (e) examples. From this perspective, it is very easy to drop the other individual out of the picture: we don't care any more exactly *who* owes the moral debt, so long as it gets paid off. (31) illustrates.

(31) a. *Deserved reciprocation*
If X does something A/U-good for Y, then X deserves to be rewarded.
b. *Deserved restitution*
If X does something A/U-bad to Y, then Y deserves to be compensated.
c. *Deserved retaliation*
If X does something A/U-bad to Y, then X deserves to be punished.

In the right-hand clauses of (31), the conclusion is only that there should be a reciprocal act that benefits the subject. The other character has disappeared; the reciprocal act may be performed by anyone (i.e. by the generic actor YA).

A questionable logical step takes us even further away from the original norms. First consider reciprocation (31a). In judging what X deserves, it is in a sense irrelevant *who* X is doing something good for. After all, rule (8) ("It's N-good to do good things for people") tells us that what's *really* important to X is that X is being N-good. So instead of characterizing X's action in terms of its A/U-goodness for Y, we can characterize it simply in terms of its N-goodness, in which Y plays no essential role. This yields (32a). The counterpart for retaliation is (32b).

(32) a. *Deserved reward*
If X does something N-good, X deserves to be rewarded for it.
b. *Deserved punishment*
If X does something N-bad, X deserves to be punished for it.

The case of restitution is slightly different. In (31b), the person who deserves restitution is Y, the person affected by the original action. So what's important to Y here is that something bad has happened to Y, and X's role is irrelevant. Thus (33) is an appropriate form.

(33) *Deserved restitution*
If something A/U-bad happens to Y, Y deserves to be compensated for it.

And there is yet another shifty step. Since one's personal normative value is a cumulative function of the N-value of one's actions, we can get from (32)–(33) to (34).

(34) a. Good people deserve to be rewarded.
 b. Bad people deserve to be punished.
 c. People that bad things happen to deserve to be compensated.

Now it's not as though the steps leading to (32)–(34) follow from any sort of formal reasoning. But intuitively they're entirely seductive.

Of course, (32)–(34) are massively counterexemplified in the world: bad things happen to people all the time with no hope of compensation, wicked people frequently do very well indeed, and all too often "Nice guys finish last." Here is the existential "problem of evil." How is it to be resolved?

Different traditions have different ways. One solution is "Virtue is its own reward" (i.e. it's A-good for you to behave in N-valued fashion), which gives up on reward coming from *outside*, and so in a way negates the spirit that leads to (34). This way of justifying virtuous behavior was alluded to in chapter 9. Christianity's solution is to put off reward and punishment until the afterlife, conveniently linking up with the strongly held belief in the survival of the soul after death (see chapter 5). Judaism tends to take the view that if something bad is happening to me now, it must be punishment for something bad I (or even my ancestors) did in the past. Hence the comedians' version of Jewish guilt: I must have done something wrong, and I'm sorry—but tell me: what was it? This also explains why many Jews lost faith during the Holocaust: nothing they or their ancestors had done could be bad enough to justify *this*.

But who is going to carry out the acts of reward and punishment? The reciprocal acts that make up for unpunished evil can't depend on people, since they are intended precisely as the way of circumventing people's injustice in the real world. Enter gods: animate moral beings who lie outside the human sphere and who take care of righting the moral scales. This puts gods in the role of protectors, beings with whom one can plead for justice and to whom one can express gratitude. Moreover, the rules of normative value dictate that one had better be nice to the gods as well, because if anyone is in a position to reward or retaliate, it's the gods. Thus the reasoning in this section leads to one of the important groundings for religion, one that is not as thoroughly explored as I think it deserves [*sic*] in recent work such as Boyer 2001 and Atran 2004 (though it is mentioned by Nemeroff and Rozin (2000) as well as, I believe, by Freud and Nietzsche).

10.5 Exchange

In the situations discussed in sections 10.2 and 10.3, reciprocation is a freely chosen act in response to a freely chosen act. Another scenario for reciprocation is *exchange*, which is basic to every sort of contract in every human society. In an exchange, the actors *agree* to do something for each other's benefit, so that the two actions are conceptually yoked more closely. In simple reciprocation, you've scratched my back, so I volunteer to scratch yours in return. In an exchange, an agreement is made: "I'll scratch your back if you'll scratch mine." "OK, let's do it." Although there may be limited cases of (semi-)agreed-upon exchange among animals (perhaps mutual grooming, mediated by nonverbal offer and uptake), the vast proliferation of exchange transactions in human cultures is unprecedented in other organisms.

The linkage between the two participants' actions identifies an exchange as a joint task in the sense of chapters 5 and 8: the actors are coactors, each performing his or her part in the task. Using the notation of section 8.8, repeated in (9) above, the overall frame for an exchange can be notated as (35).

(35) *Exchange*

$$\left[\begin{array}{l} \{\mathrm{X}, \mathrm{Y}\}\ \mathrm{ACT} \\ \mathrm{COMPOSED\text{-}OF}\ \left\{ \begin{bmatrix} \mathrm{X\ ACT}_{\mathrm{x}} \\ \mathrm{A/U\text{-}VAL}\ (\beta, \mathrm{Y}) = + \end{bmatrix}^{\beta}, \begin{bmatrix} \mathrm{Y\ ACT}_{\mathrm{y}} \\ \mathrm{A/U\text{-}VAL}\ (\gamma, \mathrm{X}) = + \end{bmatrix}^{\gamma} \right\} \\ [\mathrm{FROM}\ [\{\mathrm{X}, \mathrm{Y}\}\ \mathrm{INTEND}\ \alpha]] \end{array}\right]^{\alpha}$$

(35) is exactly the same as a simple joint task, except that it stipulates values for the participants' actions: X's part of the task benefits Y, and Y's part benefits X. We have motivated every part of this schema and related it to other forms of action and value. Yet, despite its complexity, the schema is universally part of human understanding: every human society engages in exchange, and with little apparent effort at learning. The notion of exchange presents itself as a unified gestalt.

It is useful to introduce a notational abbreviation for (35), which acknowledges its gestaltlike character in experience. From one exchange to the next, every part of the structure remains constant except for the particular actions performed by X and Y. This means that we can treat

X's and Y's actions as free variables and abbreviate the rest as a constant function *EXCH*.

(36) [X ACT$_x$] EXCH [Y ACT$_y$] (= (35))
'X performs ACT$_x$ in exchange for Y performing ACT$_y$.'

Following the discussion in section 8.8, two pairs of entailments come from the fact that exchange is a jointly intended action. The first pair is that X intends to do ACT$_x$ and that Y intends to do ACT$_y$. The second pair is that X is obligated to Y to do ACT$_x$ and that Y is obligated to X to do ACT$_y$. Because of the reciprocal obligations, neither actor is free to opt out. These entailments can be stated as (37), where the notation for obligation anticipates the treatment in chapter 11.

(37) [X ACT$_x$] EXCH [Y ACT$_y$] $\Rightarrow$
a. X INTEND [X ACT$_x$] 'X intends to do ACT$_x$.'
b. Y INTEND [Y ACT$_y$] 'Y intends to do ACT$_y$.'
c. X OB ([X ACT$_x$], TO Y) 'X is obligated to Y to do ACT$_x$.'
d. Y OB ([Y ACT$_y$], TO X) 'Y is obligated to X to do ACT$_y$.'

Like free reciprocation, exchange comes with the presumption (or default inference) that the values of the two actions are related. In the objective version, the values are equal (38a). The subjective version is more nuanced: each actor comes out ahead in terms of benefits versus costs (38b).

(38) a. *Fair exchange, objective version*
[X ACT$_x$] EXCH [Y ACT$_y$] $\Rightarrow_{\text{default}}$
A/U-VAL (X ACT$_x$) = A/U-VAL (Y ACT$_y$)
'If X performs ACT$_x$ in exchange for Y performing ACT$_y$, the A/U-values of the two acts are equal.'

b. *Fair exchange, subjective version*
[X ACT$_x$] EXCH [Y ACT$_y$] $\Rightarrow_{\text{default}}$
A/U-VAL ([Y ACT$_y$], X) + A/U-VAL ([X ACT$_x$], X) > 0
A/U-VAL ([X ACT$_x$], Y) + A/U-VAL ([Y ACT$_y$], Y) > 0
'If X performs ACT$_x$ in exchange for Y performing ACT$_y$, the value of Y's action to X outweighs the cost (negative value) of X's action to X, and similarly for Y.'

(38b) is still not complete. It omits some additional benefits to the participants. There is an A-value deriving from the experience of conducting a favorable social interaction; this is elaborated in many cultures through, say, drinking together to seal a transaction. There is also a U-value to be gained by having a trusted trading partner with whom one can anticipate future transactions, as well as a U-value that comes from being trusted. In order to gain these additional benefits, a participant may sometimes agree to a transaction in which the value of the actions exchanged would otherwise be unfavorable (Nathaniel Jackendoff, pers. comm.). These factors can be incorporated into the equation by adding a third term into the sums in (38b). This term, which encodes the value of taking part in the transaction, is indicated by a variable α that is bound to the exchange itself.

(39) *Fair exchange, subjective version (including transaction values)*

$$\begin{aligned}&[[\mathrm{X\ ACT_x}]\ \mathrm{EXCH}\ [\mathrm{Y\ ACT_y}]]^{\alpha} \Rightarrow_{\mathrm{default}}\\&\quad \mathrm{A/U\text{-}VAL}\ ([\mathrm{Y\ ACT_y}], \mathrm{X}) + \mathrm{A/U\text{-}VAL}\ ([\mathrm{X\ ACT_x}], \mathrm{X}) +\\&\qquad\qquad \mathrm{A/U\text{-}VAL}\ (\alpha, \mathrm{X}) > 0\\&\quad \mathrm{A/U\text{-}VAL}\ ([\mathrm{X\ ACT_x}], \mathrm{Y}) + \mathrm{A/U\text{-}VAL}\ ([\mathrm{Y\ ACT_y}], \mathrm{Y}) +\\&\qquad\qquad \mathrm{A/U\text{-}VAL}\ (\alpha, \mathrm{Y}) > 0\end{aligned}$$

In agreeing to an exchange, each actor is naturally trying to maximize the A/U-value of the transaction to him- or herself—following rule (10) of chapter 9: "Do what's best for you." Thus it often requires negotiation to achieve a fair exchange, in which both participants judge the exchanged acts to be of sufficient net value to themselves. Here, in the process of bargaining, is the place where microeconomics enters the picture. It is also an important point where theory of mind and Cosmides' (1989) "cheater detection" enter: one is more inclined to agree to an exchange if one believes that the other's assertions of value, agreement to joint action, and obligation are made in good faith. It may be therefore that the entailments in (39) are simply a consequence of agreeing to a joint activity: actors will not come to agreement unless each of them believes it is in his or her interest to do so. If so, (39) is redundant. Nevertheless, it is worth stating for the sake of explicitness.

Let us compare exchange with freely chosen reciprocation. Recall that freely chosen reciprocation has a normative value attached to it: "It's good to reciprocate nice things" (rule (21)). Exchange has no such normative principle, since the two actors are acting in tandem. However, exchange transactions involve a commitment to a joint task, which, as we have seen, creates mutual obligations. And failing to fulfill an obligation is bad (i.e. of negative N-value). Thus in freely chosen reciprocation the normative principle is "It's N-good to reciprocate nice things," and in exchange it amounts to "It's N-bad to default on an exchange, because it violates an obligation." This constitutes a major difference in the way reciprocation and exchange work.

The paradigm case of exchange, of course, is *trade*, or exchange of objects; a special case of trade is monetary transactions such as buying and selling. It is now easy to state these as special cases of exchange. Consider a sentence like *Bob traded Sue his goat for a coat.* Notice that this contains the telltale *for* of reciprocation, though this time the phrase following *for* denotes an object rather than an action. Nevertheless, an action is implicit: not only is Bob giving Sue his goat, but Sue is giving Bob a coat. Plugging these actions into schema (36), we get (40).

(40) Bob traded Sue his goat for a coat.
[BOB GIVE GOAT TO SUE] EXCH [SUE GIVE COAT TO BOB]

We want the understanding of this transaction to involve the R-values of the goat and the coat. Recall the intuitive definition of R-value: something has R-value to the degree that it's good to have it—that is, to the degree that having it has U-value. So we can relate R-value to U-value by the equations in (41).

(41) a. *Objective*
R-VAL (OBJECT, YA) = U-VAL ([YA HAVE OBJECT], YA)
'The R-value of an object is the U-value of having it.'
b. *Subjective*
R-VAL (OBJECT, X) = U-VAL ([X HAVE OBJECT], X)
'The R-value of an object to a person X is the U-value to X of having it.'

Now, since giving an object away changes who has it, the act of giving is of negative U-value to the giver and of positive U-value to the recipient.[6] Thus we can couch the inference rules for exchange of objects in terms of R-value as follows:

(42) [[X GIVE Z TO Y] EXCH [Y GIVE W TO X]]$^{\alpha}$ $\Rightarrow_{\text{default}}$
a. *Objective*
R-VAL (Z, YA) = R-VAL (W, YA)

[6] Notice that this entailment is not true if the entity given away is *information*. If I give you my goat, I don't have it any more; but if I tell you (give you information) that my party is on Saturday, I haven't thereby forgotten it. On the other hand, passing on information may reduce its R-value to me, for instance if I tell you where my treasure is hidden, or if I let you copy a manuscript that I hope someday to publish profitably. These considerations lie behind the motivations for spying and for copyright and patent law.

b. *Subjective*

$$\text{R-VAL}\ (\text{W}, \text{X}) - \text{R-VAL}\ (\text{Z}, \text{X}) + \text{A/U-VAL}\ (\alpha, \text{X}) > 0$$
$$\text{R-VAL}\ (\text{Z}, \text{Y}) - \text{R-VAL}\ (\text{W}, \text{Y}) + \text{A/U-VAL}\ (\alpha, \text{Y}) > 0$$

It also is easy to see how to combine exchanges of objects for actions, as in paying for services performed.

We are now in a position to go back to Fiske's fourth "elementary form of human relations," Market Pricing, in which participants exchange resources and/or labor according to negotiated agreement. In order to approach a plausible formulation of Market Pricing, let us recall the discussion of bargaining in section 5.9. There we analyzed it as conducting a cooperative task framed inside a larger presumption of competition. This means that, although the participants are taking part in a joint task, they're fundamentally adversaries, each trying to get the better of the deal. Thus the task has to be conducted in the context of a selfish or competitive stance: it's OK to get more than the other guy. Here is the selfish or competitive stance again, repeated from section 10.2.

(6) *Competition (COMP)*

For actions in the category ACT_{COMP} and individuals Y_{COMP}:

$$\text{N-VAL}\ (\begin{bmatrix}\text{YA ACT}_{\text{COMP}} \\ \text{A/U-VAL}\ (\alpha, \text{YA}) > \text{A/U-VAL}\ (\alpha, \text{Y}_{\text{COMP}})\end{bmatrix}^{\alpha}, \text{YA}) = +/0$$

'It's N-good/N-OK to act in a way that is better for yourself than for Y.'

Market Pricing amounts to specifically condoning selfishness in conducting exchanges. Thus the norm can be stated as (43) (YA_1 and YA_2 are different generic individuals).

(43) *Market Pricing (MP)*

For actions in the category ACT_{MP}:

$$\text{N-VAL}\left[(\begin{bmatrix}[[\text{YA}_1\ \text{ACT}_1]\ \text{EXCH}\ [\text{YA}_2\ \text{ACT}_2]]_{\text{MP}} \\ \text{A/U-VAL}\ (\alpha, \text{YA}_1) > \text{A/U-VAL}\ (\alpha, \text{YA}_2)\end{bmatrix}^{\alpha}, \text{YA}_1)\right] \geq 0$$

'It's OK to be selfish in exchanges.'

Notice that the outer part of (43) condones selfishness or exploitation of the other. By contrast, the inner part, the exchange, is a joint or cooperative act. Thus (43) formally instantiates the intuition that bargaining is cooperation within a larger frame of competition.

Since Market Pricing condones acting selfishly, it is best conducted between individuals for whom there is no conflicting norm to act selflessly. Thus, although one may conduct exchanges with members of one's family, it is less likely that one will insist on strict Market Pricing transactions with them than with members of another group to whom one owes no allegiance. In fact, following Jacobs's (1994) conjecture, it is plausible that Market Pricing arose in human society as a way to productively inhibit natural intergroup aggression, for mutual profit.

10.6 Linguistic Expression of Exchange of Objects and Actions

Let us conclude this chapter with a topic of hoary antiquity in linguistics (e.g. Fillmore 1965; Gruber 1965): the semantics of trading, buying, and selling. We are now in a position to be more precise about these terms than has previously been possible.

The notation for exchange so far is entirely symmetrical between the two participants. However, the linguistic expression of exchanges is asymmetrical, focusing on one side and backgrounding the other. In (40), for instance, Bob's giving of a goat to Sue is foregrounded and Sue's reciprocal giving of a coat to Bob is represented only by the *for*-phrase. In order to reflect this difference in prominence, the mapping of the exchange structure into syntax has to mark one of the actions specially. (44) indicates the difference by underlining (much as relative prominence of Stimulus and Experiencer was marked in chapter 7).[7]

(44) X trades Z (to Y) (for W).
[[$\underline{\text{X GIVE Z TO Y}}$] EXCH [Y GIVE W TO X]]

[7] As brought to my attention by (I believe) Kara Hawthorne, this account does not work for examples like (i)–(ii).

(i) The kids/Bob and Sue traded coats.

(ii) Bob traded coats with Sue.

A proper treatment of these examples calls for a more sophisticated analysis of how joint tasks are mapped into linguistic form. The syntactic form of (i)–(ii) parallels other expressions of joint tasks such as (iii)–(vi).

(iii) The kids/Bob and Sue played a duet.

(iv) The kids/Bob and Sue baked a cake together.

(v) Bob played a duet with Sue.

(vi) Bob baked a cake with Sue.

In (i), (iii), and (iv), the conjoined or plural subject maps into the set of joint actors; in (ii), (v), and (vi), the subject maps into a foregrounded member of the set of joint actors, and the object of *with* maps into the other members of the set. Examples (i)–(ii) are trickier, though, because of the bare plural *coats*. I leave this fascinating problem for future research.

Using this notation, we can distinguish the specialized verbs *sell* and *pay*. The former foregrounds the transfer of goods, and the latter foregrounds the transfer of money.

(45) a. X sells Z (to Y) (for [amount of money]).
[[X GIVE Z TO Y] EXCH [Y GIVE MONEY TO X]]
b. X pays [amount of money] (to Y) (for W).
[[X GIVE MONEY TO Y] EXCH [Y GIVE W TO X]]

The formulation in (45), however, does not allow us to bring out the parallelism between *sell* and its converse *buy*, both of which foreground the transfer of goods.

(46) a. Sue sold a goat to Bob for $500.
b. Bob bought a goat from Sue for $500.

The intuition shared in much of the literature has been that the two differ in perspective: in each case, the subject of the sentence is being thought of as the one initiating the deal, perhaps (in present terms) by making the initial offer toward joint action. This is not special to exchange verbs: a similar change in perspective appears in the *non*exchange pair *give* and *obtain*.

(47) a. Sue gave a goat to Bob.
b. Bob obtained a goat from Sue.

This difference can be expressed in terms of the macrorole tier of chapter 8: in (47a), Sue is Actor, and in (47b), the Actor is Bob. What the two have in common is the goat changing possession from Sue to Bob; following Gruber 1965 and Jackendoff 1983, we might express this as the thematic role function Z GO_{Poss} FROM X TO Y. Then the meaning of (47a,b) can be formalized as (48a,b).

(48) a. $$\left[\begin{array}{l}\text{GOAT GO}_{\text{Poss}}\text{ FROM SUE TO BOB}\\ \text{AFF SUE}\end{array}\right]$$
b. $$\left[\begin{array}{l}\text{GOAT GO}_{\text{Poss}}\text{ FROM SUE TO BOB}\\ \text{AFF BOB}\end{array}\right]$$

This change in perspective is similar to cases discussed in section 6.2, where the same thematic tier was associated with alternative macrorole tiers in a way that changed perspective.

By a parallel analysis, *sell* and *buy* might be treated as in (49).

(49) a. Sue sold a goat to Bob for $5.
$$\left[\begin{array}{l}\underline{\text{GOAT GO}_{\text{Poss}}\text{FROM SUE TO BOB}}\text{] EXCH}\\ \qquad\qquad\qquad\qquad\text{[\$5 GO}_{\text{Poss}}\text{ FROM BOB TO SUE]}\\ \text{AFF SUE}\end{array}\right]$$

b. Bob bought a goat from Sue for \$5.

$$\left[\begin{array}{ll} \underline{\text{GOAT GO}_{\text{Poss}}\text{FROM SUE TO BOB}}\text{] EXCH} & \\ & \text{[\$5 GO}_{\text{Poss}}\text{ FROM BOB TO SUE]} \\ \text{AFF BOB} & \end{array}\right]$$

Using this formal approach, we can also treat sentences that mix money and actions in exchanges, as in (50).

(50) a. Bob paid Sue \$500 to paint the house.

$$\left[\begin{array}{ll} \text{[SUE PAINT HOUSE] EXCH} & \\ & \text{[}\underline{\text{\$500 GO}_{\text{Poss}}\text{FROM BOB TO SUE}}\text{]} \\ \text{AFF BOB} & \end{array}\right]$$

b. Sue earned \$500 from Bob for painting the house.

$$\left[\begin{array}{ll} \text{[SUE PAINT HOUSE] EXCH} & \\ & \text{[}\underline{\text{\$500 GO}_{\text{Poss}}\text{FROM BOB TO SUE}}\text{]} \\ \text{AFF SUE} & \end{array}\right]$$

Next consider the noun *price*. The *price of X* is ‘the amount of money for which X can be exchanged’; that is, it foregrounds the money in a monetary transaction, relating it to the other object or action being exchanged, while leaving the actors implicit. From *price* we can build the meanings of *expensive* and *cheap*: ‘having a high/low price’. A still more complex case is the verb *owe*. Consider *Bob owes Sue \$500 for painting the house*. This expresses Bob’s obligation to Sue, where this obligation arises from a transaction in which Sue has carried out her side of the deal and Bob has not yet carried out his. In other words, this verb appeals to the entailments of EXCH as part of its meaning.

The larger point here is that all the words *trade*, *buy*, *sell*, *pay*, *earn*, *price*, *expensive*, and *owe* avail themselves of the very same conceptual structure of exchange, while foregrounding different parts and leaving implicit other different parts. This leads toward a “frame-based” theory of lexical meaning, along lines suggested by Fillmore and Atkins (1992): the notion of a transaction is a common conceptual frame that can be evoked from different perspectives and with different specializations by using different lexical items.

It is worth pointing out that this frame can be evoked in some further ways.

(51) a. Sue painted the house for \$500.
b. If you give me your goat, I’ll give you my coat.
c. You scratch my back, I’ll scratch yours.

Notice that the sentence *Sue painted the house*, by itself, contains no hint that this is for anyone's benefit (i.e. that the action is of positive U-value to anyone). However, in (51a), the addition of *for $500* brings into play the entire conceptual machinery of a transaction. Thus, from the inferences for exchanges, it is understood that some unnamed person benefits from Sue's painting the house and gives her $500 in exchange.[8] (51b) has the form of a conditional, but it is understood (presumably by implicature) as an offer to engage in a transaction. (51c) is a paratactic construction with the interpretation of a conditional (Culicover and Jackendoff 2005, chap. 13), with the same pragmatic effects as (51b). Thus all the entailments about value emerge despite the absence of overt syntactic expression.

To sum up: Chapter 9 showed that the words *good*, *should*, and *ought to* conflate a variety of conceptual predicates dealing with value that can be teased apart by careful grammatical and conceptual analysis. Similarly, this chapter has shown that the *for* of reciprocation expresses a range of value-laden relations among actions:

- A family consisting of freely undertaken reciprocation, retaliation, and restitution
- A parallel family of displays of respect including honoring, shaming, and apologizing
- The quite distinct relation of agreed-upon exchange

Using the notions of value developed in chapter 9, we have been able to formalize normative principles involving all these sorts of interaction, as well as the normative principles behind Fiske's "elementary forms of human relations," here amplified to incorporate competition. Insofar as all these interactions appear to be universal in human cultures—just differently instantiated from one culture to the next—they appear to be fundamental building blocks in the human capacity for social cognition.

[8] See Jackendoff 1990, 191–194, for an account of this effect, plus more discussion of the mapping of exchange transactions into syntax.

In addition to using verbs of exchange, there is another way in English to express the U-value of an action to a nonparticipant in the action.

(i) Sue painted the house *for Bob*. (positive U-value for Bob)

(ii) Bob's car broke down *on him*. (negative U-value for Bob)

Sentences like these were mentioned in chapter 6 in connection with the macrorole tier: the *for*-phrase is a Beneficiary and the *on*-phrase is a Patient.

Chapter 11

Rights and Obligations

11.1 Introduction

Chapters 9 and 10 dealt with normative rules such as morals, etiquette, and fairness. This chapter investigates another class of rules: rights and obligations.[1] I will be concerned especially with what might be called "social/legal/contractual" rights and obligations. Parts of the analysis will apply as well to "human rights"; other parts will not. Some of the differences between these and "moral obligations" and "moral justifications" will be mentioned in section 11.6.

As observed in chapter 5, rights and obligations are fundamental to the fabric of human social organization. Any sort of cooperation in a jointly intended task places the participants under obligation to each other to perform their respective roles (section 8.8). Ownership of an object confers on the owner (or consists of) rights to use of the object and rights to prevent others' use of it (Miller and Johnson-Laird 1976, following Snare

[1] The original version of this chapter (Jackendoff 1999) appeared in a volume in memory of my dear friend, the Irish-Canadian psychologist John Macnamara, and I was given the honor of presenting this paper as the first John Macnamara Memorial Lecture at McGill University in April 1997. The focus of John's work over three decades was an in-depth inquiry into the fundamental structure of human knowledge. Though less publicly spectacular than Chomsky's results on the nature of syntactic knowledge, John's research went beyond linguistic expression to delve into the character of meaning, thought, and reason themselves. Like Chomsky, John was asking what it is possible for a child to learn on the basis of the input in the environment—and what parts of the child's knowledge cannot be learned, but must serve as the basis for learning. Some of his results were absolutely startling, at least in the context of the sort of empiricist philosophy of psychology that prevailed at the beginning of his career and that has enjoyed a strong resurgence in these connectionist times. For those of us of a more rationalist cast, his way of chewing over issues of epistemology was a continual inspiration.

1972). Giving someone a promise places one under obligation to fulfill the promise. Conferring on someone a social status (e.g. an official title, a professional degree, or membership in an organization) grants this person certain rights and places him or her under certain obligations. Any sort of contract—including not only financial and legal contracts but also marriage in many societies—places the participants under obligation to perform certain acts. Inasmuch as the main issues addressed by a society's legal system (written or unwritten) include the privileges of ownership, the making and enforcing of contracts, and the rights and duties of officials and of citizens, it is clear that rights and obligations play a central role in understanding concepts of law.

The notions of rights and obligations, like the types of values discussed in chapter 9 and the types of reciprocation discussed in chapter 10, appear to be universal in human societies. A great deal of anthropological description is devoted to how societies differ in what rights and obligations pertain to their members, how such rights and obligations are obtained and lost, and how they are taken to be grounded in religion or government. Such descriptions invariably take the notions of right and obligation themselves for granted, not subject to discussion.[2] Yet, as I will show, these notions are remarkably complex and subtle. Thus these concepts raise interesting questions about learning and the evolution of cognition. I will turn to these questions briefly at the end.

As in the previous five chapters, I will be investigating rights and obligations within the context of a theory of conceptual structure—the level of mental structure over which inferences are defined. Like normative values, the subdepartment of conceptual structure in which the study of rights and obligations is situated is the plane of social cognition. As usual, I will be asking how people conceptualize situations in which someone can be said to have a right or an obligation. It makes little sense to ask what rights and obligations really *are*, outside of people's understanding of their social context. In other words, as discussed in chapter 5, I am

[2] To be politically correct, one might justifiably ask whether taking these notions for granted is a cultural bias on the part of anthropologists, and whether other cultures might indeed have quite different notions underlying their social organization. I don't think so: the apparent success of anthropological description—the fact that one can make sense of cultures while taking these notions for granted—suggests that there is little danger of conceptual chauvinism on this particular point of analysis. I might be wrong, of course; but I suggest that proving me wrong requires more than a blanket invocation of radical cultural relativism. See chapter 5 and Brown 1991 for discussion.

interested in a theory of the "folk theory" of social relations. Like the theory of the conceptualization of objects, space, and force in the physical domain, this forms part of the theory of human conceptualization—but one far less directly tied to perception.

11.2 The Argument Structure of Rights and Obligations

As in chapters 6–10, let us start by looking at some of the ways that rights and obligations can be expressed. We see immediately that they form a closely related pair. About the simplest way to express a right in English is with the modal verb *may*; an obligation can be expressed with the modal verb *must*.

(1) a. One may use one's possessions as one likes. (right)
b. One must pay sales tax in Pennsylvania. (obligation)

One immediate impulse for formalizing these meanings might be to take the modals to express operators (notated as *RT* and *OB*) over a proposition, as in (2). This is essentially the formalization found in von Wright's (1963) foundational work on *deontic logic*, which deals with such notions as permissions and prohibitions, the logic of *may*, *must*, *should*, and *ought*, and moral reasoning (N-values in the terms of chapter 9).

(2) a. Sue may (i.e. has a right to) leave when she wants to.
= RT (Sue leaves when she wants to)
b. Sue must (i.e. has an obligation to) leave before noon.
= OB (Sue leaves before noon)

Such a treatment, however, misses the basic point that a right or an obligation is a relation between a person and his or her action. Other readings of *may* and *must* do express propositional operators for possibility and necessity, and they lend themselves to paraphrases like (3a), whose syntactic structure reflects the semantic structure rather well. Such paraphrases are impossible with rights and obligations (3b).

(3) a. Sue may/must leave. = It is possible/necessary that Sue will leave.
b. *It is a right/obligation $\left\{\begin{array}{l}\text{that Sue (will) leave.}\\ \text{for Sue to leave.}\end{array}\right\}$

Rather, as recognized by more recent writers on deontic logic such as Forrester (1996), the proper treatment recognizes two separate arguments of these functions: the holder of the right/obligation and the situation with respect to which this person is entitled or obligated.

The first argument of these operators must be a *person*. Rocks, clouds, and computers do not have rights and obligations. Children are usually considered to have some rights, but their capacity to have obligations is age- and maturity-dependent. Animals are sometimes asserted to have rights, by construing them as semipersons; they never have obligations.[3] In modern capitalist legal thought, corporations are construed as susceptible to rights and obligations and therefore can enter into contracts; the language used to effect this construal is that corporations count as "legal persons." Thus rights and obligations are situated firmly in the plane of social cognition.

In English, the second argument of these functions must be expressed syntactically as a verb phrase (VP) whose understood subject is the holder of the obligation or right. Thus the arguments in (4a) are acceptable but those in (4b), where the VP has its own subject, are not.

(4) a. Sue has $\left\{\begin{array}{l}\text{a right}\\ \text{an obligation}\end{array}\right\}\left\{\begin{array}{l}\text{to attend the party.}\\ \text{to talk to Harold.}\end{array}\right\}$

b. *Sue has $\left\{\begin{array}{l}\text{a right}\\ \text{an obligation}\end{array}\right\}$ for $\left\{\begin{array}{l}\text{the sky to be blue.}\\ \text{Bill to leave.}\end{array}\right\}$

This VP is subject to semantic constraints similar to those for actional attitudes (chapter 8). Both *right* and *obligation* require the situation to be nonpast with respect to the time of the obligation: the VP may be present, future, or generic time.

(5) Sue has $\left\{\begin{array}{l}\text{a right}\\ \text{an obligation}\end{array}\right\}$ to leave

$\left\{\begin{array}{l}\text{ right now.}\\ \text{ tomorrow.}\\ \text{ whenever she gets annoyed.}\\ *\text{yesterday.}\end{array}\right\}\begin{array}{l}\text{(present time)}\\ \text{(future time)}\\ \text{(generic time)}\\ \text{(past time – unacceptable)}\end{array}$[4]

[3] Though I gather that in medieval Europe there were such things as trials of pigs for killing children, hence according pigs a more responsible status. Nowadays, of course, there is intense discussion about whether human fetuses have rights comparable to those of children.

[4] *Sue had a right/obligation to leave yesterday* is of course acceptable. In this case, the time for which the right/obligation is asserted is yesterday or earlier—not five minutes ago, for instance.

In the primary case of rights and obligations, the VP must express an action that the holder can carry out volitionally (6).

$$\text{(6) Sue has} \left\{\begin{array}{l}\text{a right}\\ \text{an obligation}\end{array}\right\} \text{to} \left\{\begin{array}{l}\ \ \text{leave.}\\ \ \ \text{scratch her nose.}\\ \text{*be tall.}\\ \text{*be descended from royalty.}\end{array}\right\}$$

Thus this VP expresses an intentional action in precisely the sense of chapter 8.

However, there is another case, clearest for rights: the VP may express a situation in which its understood subject receives a benefit (7).

(7) Sue has a right to be paid for her work.

(8) illustrates the difference between *right* and *obligation* in this respect. The verb *receive* does not denote a voluntary action on the part of the recipient, but the verb *accept* does. We can see this difference by applying the *What X did* test (8a). As a consequence, one can have a right to either accept or receive something, but one can have an obligation only to accept something, not to receive it.

(8) a. What Sue did was accept/*receive pay for her work.
b. Sue has a right to accept/receive pay for her work.
c. Sue has an obligation to accept/*receive pay for her work.

I will call the kind of right illustrated in (7) a *passive* right, that in (6) an *active* right.

There is in addition a very limited class of cases that might be construed as passive obligations. These are found in the context of legal punishments: *Herman must receive 40 lashes/must be banished from the kingdom/must be put to death*, where Herman cannot undertake these actions under his own volition.

I will call the person having the right or obligation the *Actor*, and the situation to which the right or obligation pertains the *Action*, with the understanding that this includes as a special case passive rights and obligations, which do not involve an Action in the standard sense. Given this much, we can formalize rights and obligations as (9), where having a right or obligation is a State, and *RT* and *OB* are two-place functions with the person and the Action as arguments.[5]

[5] This is somewhat different from the formalization in the original version of this chapter, and (I hope) somewhat simpler.

(9) a. $[_{\text{State}}\ X^{\alpha}\ \text{RT}\ [\alpha\ \text{ACT}]]$
'X has a right to do Action.'; 'X is entitled to do Action.'
b. $[_{\text{State}}\ X^{\alpha}\ \text{OB}\ [\alpha\ \text{ACT}]]$
'X has an obligation to do Action.'; 'X is obliged to do Action.'

When it is necessary to distinguish passive rights and obligations, I will use the notations *P-RT* and *P-OB*; otherwise, everything is the same.

On this analysis, the construction *have a right/have an obligation* is taken to be a light verb construction; that is, the verb *have* is a dummy and the content of the clause comes from the nominal. This construction thus is taken to parallel *have an intention/have a desire/have a tendency*, discussed in section 8.7. The only difference is that the latter cases have related verbs *intend*, *desire*, and *tend*; by contrast, it so happens that *right* and *obligation*, like *umbrage* and *opportunity*, do not.

In (9), the Action is notated as a function of one variable, its Actor; it may have further variables, irrelevant in the present context. This can be regarded as an abbreviation for the more detailed notation of the macrorole tier in chapter 8: $[_{\text{Situation}, +\text{Action}}\ \alpha\ \text{AFF}]$. The Actor position is bound to the holder of the right or obligation by the bound variable α, where the superscript α on X indicates that X binds the variable in the argument position of ACT.

The fact that rights and obligations have an Action rather than a proposition as their argument places them in the general domain of deontic logic. However, passive rights do not fall altogether comfortably into the standard deontic domain, since their arguments are not volitional actions. It is interesting therefore that the modal *may* cannot be used comfortably to express passive rights: *Sue may be paid for her work* does not paraphrase (7).[6]

The predicate OB has another argument position that plays a special role in its inferences. Suppose I have undertaken an obligation, say by promising to wash the dishes. Making a promise involves another individual to whom one has made the promise, who typically will benefit from having the promise fulfilled. (This character is not recognized, even implicitly, in the work on deontic logic with which I am familiar.) Thus the argument structure of OB should be amplified to (10).

[6] There is a similarity between a passive right to some benefit and *deserving* it, in the sense of section 10.4. In particular, most constraints on the VP complement of *deserve* are similar to those on *right* in (7). However, the parallelism is not complete. One can deserve a benefit or have a passive right to receive it. But although one can deserve punishment, there cannot be a passive right to be punished.

(10) X^{α} OB ([α ACT], TO Z)
'X's obligation to Z to do Action.'

A further constraint on the Action argument is deeply rooted in the notions of right and obligation. Essentially, a right normally concerns something one *wants* to do, while an obligation normally concerns something one *doesn't* want to do.

(11) a. Sue has a right/??an obligation to eat her ice cream sundae.
b. Sue has an obligation/??a right to scrub the toilets.

The interpretations marked *??* are sensible only if we assume Sue doesn't like the ice cream sundae and does like scrubbing toilets. (There are exceptions, however, when one has a right to do something odious or an obligation to do something pleasurable.)

I will state this intuition in terms of the *value* of the Action to the Actor: positive for a right, negative for an obligation. This is conveniently expressed in terms of the functions A/U-VAL of chapter 9.

(12) A/U-VAL ([X ACT], X) = +/−
'The A/U-value of X's action to X (Experiencer) is positive/negative.'; 'X's action is good/bad for X.'

Using this notion, we can state the constraint on rights and obligations as (13). The principles are stipulated to be defeasible to allow for cases in which other pragmatic factors intervene to create exceptions.

(13) a. X^{α} RT $[\alpha \text{ ACT}]^{\beta} \Rightarrow_{\text{default}}$ A/U-VAL (β, X) = +
b. X^{α} OB ($[\alpha \text{ ACT}]^{\beta}$, TO Z) $\Rightarrow_{\text{default}}$ A/U-VAL (β, X) = −

There are cases, such as *the right/obligation to vote*, that can be construed with either valence. In *the right to vote*, we take voting as a desirable action; in *the obligation to vote*, as somewhat burdensome. This confirms the intuitions expressed by (13). Similar effects can be discerned with the choice between *right* and *obligation* in (4a) and (5).

We may further add that the Beneficiary of an obligation normally benefits from the Action being performed.

(14) X^{α} OB ($[\alpha \text{ ACT}]^{\beta}$, TO Z) $\Rightarrow_{\text{default}}$ A/U-VAL (β, Z) = +

In addition, rights and obligations have their own values: a right is generally a good thing to have, an obligation a bad thing to have. In the terms of chapter 9, they have an R-value. We can state this as (15).

(15) a. R-VAL ([X^{α} RT [α ACT]], X) = +
b. R-VAL ([X^{α} OB ([α ACT], TO Z)], X) = −

There is an indirect connection between (13) and (15), to which we will return in section 11.6.

11.3 What One Can Do with Rights and Obligations

Next let us explore the range of things one can do with rights and obligations beyond having them.

First, one can perform the action to which the right or obligation pertains. We speak of so doing as *exercising* the right or *fulfilling* the obligation. Notice that these collocations for *right* and *obligation* involve different verbs for what (at this level of description at least) appear to be parallel actions. We will see that such differences pervade the whole range of verbs used with rights and obligations.

Second, a right or obligation can be created. Sometimes the creator of an obligation is the Actor him- or herself. For example, *promising* is (in part) creating and declaring an obligation upon oneself to perform the promised action. We speak in this case of *undertaking* the obligation. Undertaking an obligation need not be expressed with an explicit performative verb such as *promise*: it may be as simple as X saying "Will you do this for me?" and Y saying "Yes."

By contrast, though one can *declare* or *claim* one's own rights, one cannot thereby create them without the assent of other relevant parties.

A person's rights and obligations can also be created by an outside party, whom I will call the *Authority*. We speak of the Authority's *giving*, *granting*, or *conferring* rights *to* or *on* the Actor, and of *imposing* obligations. (I return to the status of the Authority in section 11.8.)

For a slightly more complex case where rights and obligations are created, consider X's *making an offer* to Y to do such-and-such. This can be construed as X's conferring the right on Y to demand (i.e. impose an obligation on) X to do such-and-such—an embedding of an obligation within a right.

Third, a right or obligation can terminate. In certain cases (as with intentions), performing the Action has this effect. For instance, handing the usher one's ticket confers on one the right to attend a performance, after which point the right ceases. Similarly, when a debt is paid, the obligation to pay it ceases. But not all rights and obligations have this property; see section 11.5.

An Actor can also cause a right to terminate by *renouncing* it. The counterpart for an obligation would be for the Actor to *reject* or possibly *renounce* it. However, renunciation of an obligation does not automatically make it terminate, even if the obligation is self-imposed; we do not think well of someone who revokes promises.

Under certain conditions, an Authority who has imposed an obligation on an Actor can *release* the Actor from the obligation, or *remove* the obligation from the Actor, in which case the Actor is *free* of it. In the case of rights granted by an Authority, we speak of the Authority's *revoking* or *taking away* these rights—in which case the Actor *loses* them.

Fourth, one can *transfer* rights from one person to another, the first party *relinquishing* them and the second *acquiring* them. The parallel case might be one person *taking on* someone else's obligations.

Fifth, in a situation of conflict between the Actor and the Authority, the Actor may *insist on* a right, which the Authority is supposed to *acknowledge* or *recognize*. Alternatively, the Actor may try to *get out of* an obligation, and the Authority may try to *hold* him or her *to* it.

Sixth, one can *infringe on* another's rights.

These situations are summarized in table 11.1.

It is beyond the scope of the present inquiry to formalize all these cases. (16) treats two basic situations, creating rights and removing them. Creating and removing obligations simply substitutes OB for RT and adds a Beneficiary. (17) analyzes offering in the same terms. (*INCH* is inchoative, or 'coming to pass'.)

Table 11.1
What one can do with rights and obligations (# indicates not necessarily felicitous)

	Right	Obligation
Performing action	exercise	fulfill
Creating		
by Actor	#declare, claim	undertake
by Authority	give, grant	impose
Voiding		
by Actor	renounce	#reject
by Authority	revoke, take away	release, remove
(effect on Actor)	lose	be free of
Transfer	transfer	take on
Conflict		
Actor	insist on	get out of
Authority	acknowledge	hold to
Other	infringe	

(16) a. Y CAUSE [INCH [X^{α} RT [α ACT]]]
'Y gives X the right to do Action.'; 'Y causes X to come to have the right to do Action.'

b. Y CAUSE [INCH NOT [X^{α} RT [α ACT]]]
'Y takes away the right to do Action from X.'; 'Y causes X to come not to have the right to do Action.'

(17) X makes an offer to Y to do such-and-such.
X^{α} CAUSE [INCH [Y^{β} RT [β CAUSE [INCH
[α OB ([α ACT], TO β)]]]]]
'X causes Y to come to have the right to cause X to come to have the obligation to Y to do Action.'

11.4 Rights and Obligations Are Not Understood Metaphorically

The use of *give*, *transfer*, and *take away* as light verbs alongside *have* suggests that a right might be conceptualized along the lines of a possession, that is, as an independent entity that one may have, give, or take away. *Infringe* has associations of treading on another's territory, a variation on possession. The language associated with obligations has more inconsistent and opaque connotations. *Undertake*, *impose*, *remove*, and *take on* seem to image the obligation as a burden to be borne, as does the phrase *under (the weight of) obligation(s)*. On the other hand, *release*, *get out of*, and possibly *hold to* suggest the obligation is imaged as a constraining force. In particular, the notion of an obligation as a constraint relates it to force-dynamic expressions (Talmy 1988): it is a social force that affects one's course of action. The expression *fulfilling an obligation*, through its association with *fill*, might suggest an image of the obligation as a container. The almost synonymous phrase *meeting an obligation* carries overtones of yet another image, one whose character is difficult to pinpoint.

Let us consider these "images" a little further. A theorist in the vein of Lakoff and Johnson (1980) would claim that rights and obligations are understood "metaphorically" and that they derive their conceptual properties from another domain, called the "source domain." According to Lakoff and Johnson's methodology, the evidence for identifying the source domain comes precisely from the collocations in which the words in question appear. In this particular case, we would be inclined to claim that rights and obligations are understood metaphorically in terms of *different* source domains—rights as possessions, obligations as burdens or constraints. Yet, as we have already seen to some extent, and as I will continue to document, rights and obligations are near-twin concepts, with altogether parallel logic. So there is something a bit suspicious about the metaphor view.

An alternative view is that rights and obligations have their *own* logic. This logic is shared superficially with possessions and burdens, but it is close enough to draw an associative connection. In choosing verbs to express what one can do with rights, the language is swayed toward verbs of possession because rights, like possessions, generally are of positive value; verbs relating to obligations are swayed toward verbs of physical burden and constraint because obligations, like burdens and constraints, generally are of negative value. On this view, one does not understand rights and obligations metaphorically in terms of possessions and burdens or constraints. Rather, because of what one understands about rights and obligations, one chooses verbal collocations in a motivated fashion. While acknowledging the insights that Lakoff and Johnson seek to express, this view turns the notion of metaphorical understanding on its head: it is because rights and obligations are understood as they are that the metaphorical connection is possible—not the other way about. (See Murphy 1996; Jackendoff 1992a, chap. 3; and Jackendoff and Aaron 1991 for more detailed discussion of Lakoff and Johnson's approach.)

Looking a little more deeply, what sort of conceptual entities might rights and obligations be? In the formalization in (9) and (16), having a right is being in a certain social state. Yet the grammar makes the right look like an independent abstract entity, which someone may have, may be given, or may have taken away, and toward which someone may express attitudes such as insistence or acknowledgment. A more or less standard philosophical approach to rights and obligations, observing that they involve clausal complements, might suggest that they are propositional attitudes, like beliefs and desires.

There are two reasons why I reject this view. First, there are good arguments against the standard treatment of propositional attitudes (see chapter 8). They are attitudes not toward propositions, but toward conceptualized situations and actions. Moreover, it is incorrect to reify beliefs and desires as abstract objects; rather, the words *belief* and *desire* are simply nominals of the verbs *believe* and *desire*. Thus the fundamental semantic unit for considering propositional attitudes is a state of someone believing or intending something. This concurs with the treatment of rights and obligations here. In particular, the fundamental predicate is *having* a right or obligation, and the arguments of rights and obligations are actions rather than propositions.

But even if one were to accept the standard treatment of beliefs and desires, there is a more important reason to reject it for rights and obligations: rights and obligations are emphatically not attitudes. Beliefs and intentions are conceptualized as being "in an individual's mind"; having a belief or intention is being in a subjective mental state. By contrast, having a right or obligation is being in a certain objective social situation. To make this contrast clearer, notice that *Bill's belief/desire is in his mind* is a sort of tautology. By contrast, *Bill's right/obligation is in his mind* attributes to Bill some kind of delusion about his social relations. In fact, understanding someone's rights and obligations requires no understanding of that person's mind at all. Here rights and obligations differ sharply from the more standardly studied species of deontic concept, moral/ethical understanding. As seen in chapter 9, moral understanding can depend heavily on one's understanding of others' minds and others' subjective values (Hoffman 1987; Goldman 1993); this dimension is absent from rights and obligations.

11.5 Existentially versus Universally Quantified Rights and Obligations

Section 11.3 spoke of rights and obligations that terminate by virtue of exercising or fulfilling them. To repeat, handing the usher a ticket gives one the right to attend the performance—once; then one has the right no longer. Paying a debt erases one's obligation to pay it; returning a borrowed item erases one's obligation to return it. On the other hand, not all rights and obligations are like this. Showing the usher one's year-long pass to the theater gives one the right to enter the theater, but one retains the right for future occasions. Similarly, one's obligation to obey a police officer does not end when one obeys an officer once: this obligation persists.

We must distinguish, then, between two types of rights and obligations: those that pertain to exactly one action and those that pertain to all actions of a given type. The former are exercised or fulfilled by an appropriate action's taking place, at which point they terminate. We could think of these as *existential*, in the sense that if there comes to exist an Action that satisfies the argument of the right or obligation, then the right or obligation ceases to exist. The latter, by contrast, are *universal*: they pertain to every Action of the appropriate type. (Von Wright calls these varieties "particular" and "general," respectively; the legal tradition uses the terms "*in personam*" and "*in rem*," respectively.)

This distinction could be encoded by a subscript on the functions RT and OB, for example RT_{Ex} and OB_{Ex} versus RT_{Un} and OB_{Un}. Inference rule (18) then pertains only to the existential variety. (For convenience, I omit the Beneficiary of OB.)

(18) X^{α} RT_{Ex}/OB_{Ex} [α ACT] at time t_1, and
X ACT at time t_2, where $t_2 > t_1$,
$\Rightarrow$ NOT [X^{α} RT_{Ex}/OB_{Ex} [α ACT]] at time t_3, where $t_3 > t_2$

It is important to our story that this inference rule involves a succession of times—that is, the principles of rights and obligations require a dynamic logic. (This temporal dependency is not present in the formalism of von Wright and his successors, even if they acknowledge it informally.)

As mentioned in chapter 8, various other concepts have this property. For instance, an intention to perform an action is fulfilled—and thereby terminates—when the action is performed. Parallel effects obtain with bodily sensations such as hunger, thirst, and some itches. So this property of rights and obligations is actually more broadly attested among our concepts.

The situation is actually even more complex. One has the right to vote in every election; that is, exercising the right does not eliminate it. On the other hand, one has the right to vote only once in each election; having voted eliminates the right *till the next election*. So this right has a mixed flavor, partly universal and partly existential. This suggests that the simple subscripts *Ex* and *Un* are not subtle enough to capture the range of possibilities; there is more internal structure to be teased out.

11.6 The Fundamental Principles of Rights and Obligations: Consequences of Noncompliance

So far I have not done much to distinguish social/legal/contractual rights and obligations from other kinds of deontic operators expressed by modals such as *may*, *should*, *ought to*, and *must*, in particular from “moral obligations” and “human rights,” which turn out to be a bit different in character. Now I turn to what gives social/legal/contractual rights and obligations their distinctive flavor: the consequences of noncompliance.

What happens if one fails to fulfill an obligation? Very simply, one runs the risk of getting in trouble. Suppose I have undertaken an obligation, say by promising you (the Beneficiary) that I will wash the dishes. Now suppose I do not wash the dishes within a reasonable amount of time. (What counts as a reasonable amount of time is a delicate matter.) Then you have the (existential) right to impose some sort of punishment on me. You do not necessarily exercise this right, but we clearly understand that this right exists.

What does it mean to impose punishment? Following section 10.2, if Z punishes X, Z performs some action with negative A/U-value to X, *in return for* some previous action on the part of X with negative value to Z. (*Reward* is the same, except with positive values.)

This is enough information to enable us to state a preliminary version of the inference rule for nonfulfillment of obligations; again it is stated dynamically (i.e. with a time-dependence).

(19) *Fundamental principle of obligations*

X^{α} OB ([α ACT_1], TO Z) at t_1 and
NOT [X ACT_1] in period from t_1 to t_2

$$\Rightarrow Z^{\beta}\ \mathrm{RT}_{\mathrm{Ex}} \begin{bmatrix} \beta\ \mathrm{ACT}_2 \\ [\mathrm{RECIP}\ [\mathrm{NOT}\ [\mathrm{X}\ \mathrm{ACT}_1]]] \\ [\mathrm{A/U\text{-}VAL}\ (\gamma, \mathrm{X}) = -] \end{bmatrix}^{\gamma} \text{ at } t_2$$

'If at t_1 X has an obligation to Z to perform ACT_1, and X does not perform it by t_2, then Z has the right at t_2 to do something of negative A/U-value to X in retaliation for nonperformance.'

In the consequent of (19), ACT_2 is the action that Z has the right to carry out; the RECIP operator (section 9.5) says that this action is in return for X's nonperformance; the value expression says that ACT_2 is of negative A/U-value to X. Z may or may not choose to retaliate, but (19) gives Z the right to do so.

Rule (19) says nothing about the appropriate time interval to wait for compliance. More important, it says nothing about what kind of retaliatory act is appropriate—only that it should be something that the Actor won't like. Many such actions, especially for culturally loaded obligations, are prescribed in a culture's stock of customs and oral or written law. In particular, there are large classes of obligations, including such things as debts, for which the appropriate action in the face of noncompliance is to call in the authorities to determine the appropriate punishment. In turn, social norms may dictate that the authorities have an obligation to the Beneficiary to mete out the appropriate punishment—one of the foundations of legal theory (Stone 1968). I return to the notion of "the authorities" in section 11.8.

Turning to rights, we find a related situation. Suppose I have given the usher my ticket. This gives me a right to enter the theater to see the performance. Now suppose someone tries to prevent me from going in, saying "You may not go in!" I am thereby entitled to take action against this person: my right has been violated. Notice, by contrast, if I have tried to go in without presenting a ticket, I'm not entitled to express disapproval, since it's the rights of the theater (as corporate body), not mine, that are being violated.

This scenario leads us to the counterpart of rule (19) for rights. If I have a right, I may or may not choose to exercise it. But if I attempt to exercise it, and some other party tries to prevent me, I then have the (existential) right to exact punishment on that person. (20) formalizes this intuition.

(20) *Fundamental principle of rights*

$$\begin{array}{l} \mathrm{X}^{\alpha}\ \mathrm{RT}\ [\alpha\ \mathrm{ACT}_1]\ \text{at}\ \mathrm{t}_1\ \text{and} \\ \mathrm{Z\ CAUSE\ NOT}\ [\mathrm{X\ ACT}_1]\ \text{at}\ \mathrm{t}_1 \\ \quad \Rightarrow \mathrm{X}^{\beta}\ \mathrm{RT}_{\mathrm{Ex}} \begin{bmatrix} \beta\ \mathrm{ACT}_2 \\ [\mathrm{RECIP}\ [\mathrm{Z\ CAUSE\ NOT}\ [\mathrm{X\ ACT}_1]]] \\ [\mathrm{A/U\text{-}VAL}\ (\gamma, \mathrm{Z}) = -] \end{bmatrix}^{\gamma} \end{array}$$

'If X has a right to perform some action and Z prevents it, then X has the right to do something of negative A/U-value to Z in retaliation.'

Again, there are many cultural customs and norms concerning what kind of retaliatory ACT_2 is appropriate to what kind of ACT_1, and under what sort of relationship between X (the Actor) and Z (the Right-Violator). In particular, just as in the case of unfulfilled obligations, X's right in many cases will consist of a right to go to the authorities to demand retaliation or restitution. For instance, if the usher doesn't let me into the theater despite my having presented my ticket, the appropriate action, after due remonstration, is to go to the manager, and if that doesn't work, to the police.

In particular, it appears that the basic forms of *laws* are mostly variations on inference rules (19) and (20): "Anyone of such-and-such a category is *obligated* to the group/state to perform the following actions, and if not, the following punishment is to be imposed." "Anyone of such-and-such a category is granted by the group/state the *right* to perform the following actions, and anyone violating that right is subject to the following punishment." In turn, the punishment may be an obligation to pay a fine or perform certain actions; it may be a passive obligation to endure privation; or it may be the loss of rights. Another major category of laws establishes institutions, that is, stipulates categories of individuals and groups (e.g. the legislature, the courts, marriages). But the point of doing so is precisely to establish the rights and obligations of such individuals and groups.

Let's look at the role of reciprocation in (19)–(20), recalling the discussion in section 10.2. The default inference is that the value of the reciprocal action is equal to that of the original action. (21) repeats the principle stated in section 10.2.

(21) *Principle of reciprocation*

$$\begin{bmatrix} \text{Y ACT}_2 \\ \text{RECIP [X ACT}_1]^{\alpha} \end{bmatrix}^{\beta} \Rightarrow_{\text{default}} \text{A/U-VAL}\ (\beta, \text{X}) = \text{A/U-VAL}\ (\alpha, \text{Y})$$

'When Y acts in return for X's acting, Y's act is as good/bad for X as X's act is for Y.'

Thus (21) helps guide what actions are appropriate in retaliation for breaking obligations and violating rights in (19) and (20): "The punishment fits the crime."

Combining (21) with (19) and (20) gives us an inferential link between (13) and (15), that is, between the value of rights and obligations and the value of the actions to which they pertain. Consider first obligation. An action that one is obligated to perform has a cost, and *not* performing the action risks a cost (if the Beneficiary exercises the right of retaliation). Hence having an obligation is basically a lose-lose situation; that is, the obligation itself is of negative value. Similarly, an act that one has a right to perform is of positive value, and being prevented from exercising the right grants one a right of retribution or restitution, also a positive value. Hence having a right is a win-win situation, that is, a benefit. So some of the pieces of the logic of rights and obligations begin to hang together.

In addition to (19), the nonfulfillment of an obligation has a broader consequence. Roughly speaking, *everyone*—not just the Beneficiary—is justified in criticizing the Actor (or thinking less of the Actor) for failing to fulfill the obligation. The threat of such criticism seems to me to constitute the *moral* dimension of an obligation. Here I would be reluctant to use the term "has a right" to describe what everyone may do to the Actor; something like "is morally justified" seems more appropriate. Using the notion of N-value from chapter 9, this principle can be stated as (22).

(22) $$\text{N-VAL}\ (\begin{bmatrix}\text{X}^{\alpha}\ \text{OB}\ ([\alpha\ \text{ACT}_1],\ \text{TO Z})\ \text{at}\ \text{t}_1\ \text{AND} \\ \text{NOT}\ [\text{X ACT}_1]\ \text{in period from}\ \text{t}_1\ \text{to}\ \text{t}_2\end{bmatrix}) = -$$

'It's bad not to fulfill your obligations.'

In turn, by the principles of chapter 9, if X commits an act of negative N-value, X's esteem (E-value) in the eyes of the community goes down.

These two inferences, the one social/contractual, the other ethical, seem both to be involved in social/contractual obligations such as promises and debts. The kinds of things we might call "moral obligations," such as the obligation to preserve the environment, seem to me to invoke only N-value.[7] The two inferences are not altogether parallel in structure. In particular, only the *Beneficiary* of an obligation has a right of retribution, whereas *everyone* is morally justified in disapproving of the offender.

As in the case of obligations, there is also a moral dimension to violation of a right: everyone is morally justified in criticizing (or thinking less of) the person who violates someone's rights. Again, only the Actor has the right of *retribution*, but everyone is morally justified in criticizing the offender.[8]

The fundamental principles of obligations and rights ((19) and (20), respectively), though nearly symmetrical, have one important difference. In the case of obligations, there is a specific Beneficiary of the obligation. As seen in (19), this individual acquires the right of retaliation if the obliga-

[7] What I am calling here the moral dimension is the main one considered by Conison (1997), who draws the inference that everyone is entitled to sanction someone who breaks a promise. Von Wright (1963, 12) however, says, "'By definition', one could say, promises ought to be kept. But this is only one aspect, beside others, of the obligation in question.... To try to explain the obligation to keep promises, for example, in terms of the 'normative pressure' of customs seems utterly out of place." This is the intuition I am trying to capture here.

[8] Principle (20) also does not apply comfortably to what many take to be the most important kinds of rights, namely universal human rights. There are some differences between these and the more mundane type of contractual rights discussed here. Unlike issues of contractual rights, issues of human rights typically arise in the context of *claiming* rights that are not acknowledged by governmental authority. The Declaration of Independence asserts that such rights are granted by higher (divine or natural) authority (as in "...endowed by their Creator with certain unalienable rights..."); in the past century, they have been to some extent acknowledged through international agreements such as the United Nations Universal Declaration of Human Rights. Conflicts concerning human rights arise from a failure of mutual belief in the existence of the rights in question, as well as from questionable effectiveness in enforcement on the part of higher authorities such as governments.

tion is not met. In the case of rights, there is no such specifically identified individual. *Anybody* who tries to prevent one from exercising a right is a potential target for justified retaliation.

This difference, however, conceals a deeper similarity: the individual entitled to retaliation is always the one for whom the potential Action is a benefit. In the case of a right, the Actor is the potential Beneficiary and therefore receives the right of retaliation for interference with receiving the benefit. In the case of an obligation, the Action is of negative value to the Actor; the Action is being performed to benefit another individual. It is that other individual, then, who receives the right of retaliation.

Notice too that this parallelism extends to the moral dimension. Everyone is morally justified in criticizing the person who prevents the intended Beneficiary from receiving the benefit: the Actor in the case of an obligation, the person who interferes with the Actor in the case of a right.

My sense is that inference rules (19) and (20) are the central principles that make rights and obligations what they are. By contrast, a moral/ethical principle takes the form "One should/should not do such-and-such"; that is, it is a similar relation of an individual to an action in terms of N-value. But it does not carry inferences concerning rights of retaliation. Rather, it carries only the normative dimension, the fact that everyone is morally justified in criticizing the Actor for noncompliance. It is the need to regulate modes of retaliation in response to violations of social/contractual obligations and rights that leads to the development of legal and judicial systems in a society. So these inference rules lie at the foundation of social/cultural cognition, as well as at the root of legal systems.[9]

11.7 Reciprocal Rights and Obligations

To amplify the symmetries we have been observing: It is often felt that there is a sort of reciprocity between particular rights and obligations. In particular, the Beneficiary of an obligation is felt to have a "passive right" to receive the benefit. We can state this sense as inference rule (23).

(23) X^{α} OB ([α ACT], TO Z) $\Rightarrow$ Z P-RT [X ACT]
'If X has an obligation to Z to do something, then Z has a passive right to have that action performed.'

[9] As mentioned in section 5.10, in certain religious traditions such as Judaism, moral/ethical strictures are taken to be obligations to a deity, and the deity acquires the right of retaliation. I take this to be a cultural construct whose purpose is to sharpen ethical norms, giving them the same "objective" status as laws.

The consequent of (23) is a passive right (P-RT) because Z is not the Actor of the Action. However, since X's obligation is to act for the benefit of Z, X's action satisfies the conditions for the argument of a passive right, as illustrated in (7) and (8b) above (*Sue has a right to be paid for her work*).

Now let us apply inference rules (19) and (20) to both sides of (23). When we apply (19) to the left-hand side (X's obligation), the inference is that Z is entitled to retaliate against X if X does not perform the Action. When we apply (20) to the right-hand side (Z's passive right), the inference is that Z is entitled to retaliate against anyone who prevents Z from receiving the benefit—in particular against X if X does not perform the action. So both sides of (23) lead to similar inferences if X does not comply.

A right imposes a similar sort of reciprocal obligation: the obligation falls on *everyone* not to infringe on someone's rights. In particular, the individual who grants the right is bound to respect it. The general rule can be stated as (24). (The sense of "everyone" is conveyed by the generic individual YA introduced in chapter 6.)

(24) X^{α} RT [α ACT] ⇒
YA^{β} OB ([NOT [β CAUSE [X NOT ACT]], TO X])
'If X has a right to perform some action, everyone (generic individual) is obligated to X not to prevent X from acting.'

Notice that the obligation in (24) is to *not* perform an action (or to *refrain* from performing it)—a sort of counterpart of a passive right. Again, the outcome of someone infringing on a right, as entailed by rule (20), is the same as the outcome of someone failing to meet the obligation not to prevent exercise of a right, entailed by rule (24). So perhaps (24) is logically unnecessary. Nevertheless, something like it is often stated explicitly (e.g. in Stone 1968), so I include it for completeness.

11.8 Authority

Consider who can impose an obligation on you. As pointed out in section 11.3, the simplest case is a self-imposed obligation such as a promise. Other things being equal, you are free to make whatever promises you wish. But no one else can impose an obligation on you unless particular conditions obtain. For example, if a random person declares, "I hereby oblige you to wash my feet," you are justifiably offended and baffled. The felicity conditions for such a performative speech act are not met, just as if a random person were to declare, "I hereby name you Fuzzy-Wuzzy."

One felicitous condition under which an obligation can be imposed by someone else comes about if you have granted the other person the right to impose it, either by making an offer or by making an agreement or contract. You are perfectly free to grant such a right.

However, there is another situation in which someone can impose obligations on you: when that person has authority over you through position in the social hierarchy. A society presents many authority relationships, such as parent to child, boss to worker, sergeant to private. (However, not every dominance relationship is an authority relationship; for instance, the dominance hierarchy among siblings by age may not confer authority, depending on the culture.) The authority relationship grants the authority the right to impose obligations on the subordinate. If, for a first approximation, we encode the authority relationship as (25), then (26) expresses the authority's right to create obligations for subordinates.

(25) Z AUTH X
'Z has authority over X.'

(26) Z AUTH X $\Rightarrow$
Z^{α} RT [α CAUSE [INCH [X^{β} OB ([β ACT], TO α)]]]
'If Z has authority over X, then Z has the right to cause X to come to have obligations to Z.'

The language used to express what one can do with authority is virtually identical to that for rights listed in table 11.1. We speak of the *exercise* of rights and of authority; a higher authority can *give* or *grant* rights or authority—or *revoke* them or *take* them *away*. One can *renounce* rights or authority that one currently *holds*; or one can through malfeasance *lose* rights or authority. One can *insist* on one's own rights or authority; one can *acknowledge* someone else's. On the other hand, there is more complexity in authority, since one can *resist* another's authority, but the phrase *resisting another's right* makes little sense.

Like all obligations, those imposed by an Authority must have a Beneficiary. (26) encodes the typical case in which the Beneficiary is the Authority him- or herself (this is notated by *TO* α in the third argument position of OB, bound to Z). Thus in case of noncompliance, the authority also has the right of punishment. Other Beneficiaries are possible, for instance when a judge obliges a divorced parent to pay child support to the ex-spouse. In such a case, the ex-spouse's right of retaliation for noncompliance is typically determined by the judge as well. That is, if an authority imposes an obligation on an Actor, with a third individual as Beneficiary, the Authority retains the right to punish the Actor for noncompliance, with or without appeal from the Beneficiary. So the logic becomes still more complex; I will not attempt to formalize it here.

Even with these added complexities, the account in (26) is still missing an important caveat, in that it is necessary to recognize limitations of authority. For instance, in our society, we believe that a boss does not have the right to oblige an employee to engage in sexual activity. A more adequate formalization relativizes authority to a particular class of actions, as in (27).

(27) Z AUTH (X^{β}, β ACT_A)
‘Z has authority over X with respect to actions of type A.’

The appropriately revised form of (26) is (28) (where I also add the Authority’s right to confer rights).

(28) Z AUTH (X^{β}, β ACT_A) ⇒
Z^{α} RT [α CAUSE [INCH [X^{β} RT/OB ([β ACT_A], TO α)]]]
‘If Z has authority over X with respect to actions of type A, then Z has the right to cause X have rights or obligations to Z to perform actions of type A.’

(28) leaves about the right loopholes for social negotiation (and conflict): over exactly what actions can a given authority impose obligations? And how are those decided? These are issues with which every society must grapple, and which form an important segment of legal systems.

How does one obtain authority? One way is to be granted it by a higher authority, who is then said to be *delegating* authority. But this leaves open who grants authority at the top of the pyramid. This problem of the “apex norm” (Stone 1968, following Hans Kelsen) lies at the root of a society’s conception of itself. Here are four possible solutions; perhaps there are others:

- Despotism, where the ultimate authority simply asserts authority without recourse and maintains it through the exercise of force
- Supernatural authority such as the “divine right of kings,” in which the top-ranked person is said to be granted authority by a deity, whose rights in turn require no justification
- Representative government, in which authority is taken to arise from the “consent of the governed,” that is, from a joint action or joint commitment on the part of the group

- The “natural” authority of parents over children, which seems to need no justification and which is recognized in all societies

11.9 Where Does It Come From?

I have surely left many subtleties still untouched—and some major points as well, such as how to reason about conflicting rights and obligations, and how to characterize rights as legitimate or illegitimate. Nevertheless, let us now step back a bit.

We have found that the concepts of right and obligation are quite abstract, not linked to perception of the physical world except very indirectly. The analysis here suggests that, like values, their content lies entirely in the inferences that can be drawn from them. They are, as it were, part of an elaborate social accounting system for keeping track of the implications of an individual’s actions with respect to others, a system rooted ultimately in the notion of value. However, unlike the sorts of value discussed in chapter 9, rights and obligations do not come in both subjective and objective flavors: rights and obligations are conceptualized as objective abstract entities in the social domain.

The central elements of the accounting system are the principles (19) and (20), which define obligations and rights through their inferences, which in turn invoke further rights. These principles depend on the notion of linking two actions as reciprocation measured in terms of costs and benefits, and through the principles of reciprocation they are connected to notions of fairness or justice. The fundamental notion of authority is in turn defined largely in terms of rights. In short, many of the conceptual foundations of social organization either depend on an understanding of rights and obligations, or else are developed to justify the assertion of particular rights and obligations.

To function in a society, then, it is essential that a person intuitively grasp the concepts of right and obligation. Indeed, most of the discussion here has consisted of pointing out intuitions that all of us share. So the question arises, how do people acquire these concepts? As Macnamara (1991) puts it, how does one gain entry to a system of interrelated terms and ideas, if they cannot be defined in terms of some other system?

There is no question that people must learn the particular network of rights and obligations inculcated (or presupposed) by their society: who has an obligation to whom by virtue of what roles they have, who has a right to impose obligations and grant rights over what actions, what retaliation is appropriate for failure to meet what obligation, and so forth.

This must by all means be a major part of cultural learning, along with culture-specific normative principles. But, returning to a point stressed in chapter 5, it is far less clear that people must learn that there *are such things* as rights and obligations. As far as I know, every culture shares these concepts. They seem to be building blocks as fundamental to understanding the social world as force is to understanding the physical world. (This point is also made by Forrester (1996).)

Moreover, the inferential patterns of rights and obligations have no analogue in the physical (or sensorimotor) domain, such that there could be a progression in learning as suggested by Piaget, or a learning through metaphor as suggested by Lakoff and his colleagues. The latter possibility was rejected in section 11.4, even before we approached the complexities of retaliation and exchange, for which a physicalistic metaphor is still more far-fetched. In fact, if anything, the tendency often goes the other way: people attempt to understand the physical world by anthropomorphizing it into a metaphorical social world full of wills and desires, and often by passing off the volition to supernatural social beings such as deities.

It seems to me, therefore, that an important question for research into social cognition is how the child learns the concepts of right and obligation—if they are learned at all. The latter possibility, not to be discounted, is that these concepts are largely if not entirely innate, a specialized "way of thinking" wired into the brain by the human genome. Such an account would certainly explain the cultural universality of these concepts: they would form a preestablished species-wide skeleton of social understanding over which each particular culture builds its own flesh. Under this hypothesis, the child learning a culture would come to the task predisposed to interpret the social world in terms of rights and obligations, among other things. If there is an identifiable developmental stage where such concepts become available, relatively uniform across cultures, this might well be interpreted as evidence of biological maturation of the brain. If so, the argument would be parallel to the arguments for the biologically based language capacity that makes language acquisition possible (Chomsky 1965; Lenneberg 1967; Pinker 1994; Jackendoff 2002a; see also chapter 2).

I am not aware of any research that bears directly on the acquisition of these precise concepts. However, suggestive evidence appears in the experimental work of Piaget (1932), who discusses the development from age 6 to 11 of the child's understanding of related deontic concepts such as the rules of games and of moral concepts such as prohibition and fairness, as well as their relation to authority. In particular, since the rules of games have some of the same "objective" ontological status as obligations, they should provide suggestive evidence.

Looking earlier in child development, work by Cummins (1996a,b) and Harris and Núñez (1997) investigates 3- and 4-year-old children's understanding of the related deontic concepts of prohibition (denial of a right), permission (granting of a right), and reciprocal exchange. They find a degree of understanding of these concepts more sophisticated than one might have expected from Piaget's research, in fact an understanding more reliable than with equally complex propositional statements. For instance, Cummins compares 3- and 4-year-olds' understanding of a prohibition such as "All squeaky mice have to stay in the house" with that of a declarative proposition such as "All the squeaky mice are in the house." The children's understanding is tested by asking them which (toy) mice they have to check to determine whether the order has been carried out or the proposition is true. The question is phrased in such a way that the very same mice must be checked under both conditions. Cummins finds that the children answer the question about the prohibition much more reliably than the question about the proposition. (Notice also that this is not exclusively ego-centered deontic understanding, as might be acquired through a child's experience with parents' orders: the task involves checking whether other individuals have obeyed someone else's order.) Cummins concludes that this aspect of deontic understanding is in place early on in development. This is not a test of the entire logic of rights and obligations, but at least an important part.

If the logic of rights and obligations were part of the human endowment, it would have likely emerged from some evolutionary antecedent. So some precursors might be expected in the social behavior of primates. Chapter 10 alluded to reciprocal altruism and aggression in primates as precursors of the human versions; whether there are precursors of the more elaborate notions of rights and obligations is an open question.

Three directions suggest themselves for investigating the issue further. First, crosslinguistic and crosscultural work on the language and understanding of rights and obligations would add a great deal to the analysis. Second, I have barely scratched the surface of the formal detail of rights and obligations, their dynamic functioning in social reasoning, and their relationships with other social concepts (including authority, law, and moral thought). To figure out exactly what the child has to learn—and what the child *can* learn—it is crucial to pull these concepts apart further into their components and to see what external evidence could lead to acquiring such components. A third direction is to use formal analysis of the sort developed in this chapter and the previous ones to help guide further research in anthropology, primatology, and especially child development on this topic that is so vital to our social existence. It is my hope that researchers better versed than I in these disciplines will be stimulated by this discussion to undertake such a challenge.

Chapter 12
Trumpets and Drums

12.1 Methodology in Studying Social Cognition and Theory of Mind

Stepping back from the mass of detail, where are we?

Chapter 5 proposed to investigate human social and cultural cognition along lines parallel to theoretical linguistics. The essence of social cognition lies in the interaction of abstract (i.e. nonperceptual) concepts such as beliefs and intentions, values, reputations, rights and obligations, and group membership. All these are in turn rooted in the concept of a *person*, itself an abstract concept, in that a person is conceptualized as something over and above his or her body. Social/cultural behavior therefore must involve linking these abstract concepts to perceptible objects and actions, which are thereby accorded a social meaning.

Ideally, in the end we would like to know how the brain accomplishes social cognition. However, at the moment the tools of neuroscience can only give us relatively crude accounts of, for instance, where in the brain some particular aspect of processing takes place, the fact that certain neurons fire in response to very particular stimuli, the timing of certain signals of brain activity, the cognitive consequences of various lesions, and the overall character of massively parallel interaction at every scale from the smallest neural assemblies to the entire nervous system.

Such results tell us little about important details of social/cultural knowledge: exactly what constitutes theory of mind, how an "existential" obligation differs from a "universal" obligation, how an "objective" value differs from a "subjective" value, how free-choice reciprocation differs from agreed-upon exchange, how jointly intended action differs from individual intended action, how group membership interacts with value, and on and on. Much less does neuroscience tell us how all these conceptual structures affect the formulation of behavior, except perhaps in the most general terms.

For the moment, then, if we wish to study the details of social/cultural cognition, it is necessary to fall back on older formal tools of logic and theoretical linguistics, adapting them to the new context. This is what I have attempted here. I find it faintly disturbing that such formal analysis should require defense against wet neuroscience and brain imaging. In order to understand the brain, we need all the tools we can get, and in principle the approaches should complement rather compete with each other. Yet it is customary for neuroscientists to disregard the results of theoretical linguistics, and I have been told that they are explicitly taught to do so. Moreover, after over 20 years of research, the predominant paradigm in computational modeling of neural activity, connectionism, is incapable of representing the most basic combinatorial properties of language, and when pressed, persists in denying that such properties are significant (see Bybee and McClelland 2005 for such denials).

This situation exists partly because the highest-profile neuroscience looks at vision, not language. One can only study so many things in one's lifetime, especially with grants and tenure at stake. But, as I am the first to admit, the situation is also a result of a pervasively cavalier attitude within theoretical linguistics toward more general issues of cognitive neuroscience, including even language processing. As documented in chapter 2 (see also Jackendoff, forthcoming), this attitude grows in part out of theoretical misconceptions within the field itself. Here and elsewhere, I have attempted to reorient theoretical linguistics in such a way that productive dialogue is possible (see chapter 2, summarizing Jackendoff 2002a and Culicover and Jackendoff 2005). I recognize that one person's efforts to move several large intellectual communities toward each other after decades of mistrust may be a quixotic undertaking, but it is all that one person can do.

Studying social cognition in this context adds a further layer of tension. I have chosen to pursue the inquiry within the context of Conceptual Semantics, an approach to meaning and conceptualization that is not only foreign to the methodology of most cognitive neuroscience, but also depends on a theory of language foreign to predominant strains of linguistics, logic, and philosophy. This approach assumes that concepts can be studied (which, surprisingly, makes many linguists unhappy) and that concepts are instantiated in the brain (which makes neuroscientists happy but many philosophers unhappy). It argues that concepts have a formal structure (which leaves many neuroscientists cold) and that the concepts expressed by words have combinatorial structure (which makes many philosophers and logicians unhappy). It further argues that the combinatoriality of concepts expressed by phrases is not purely a consequence of syntactic combinatoriality (which makes many linguists, logicians, and philosophers unhappy) and that the structure of concepts is not due entirely to the structure of the world (which makes psychologists happy but many philosophers furious). The arguments for this approach were sketched very briefly in section 6.1 and have been addressed in painstaking (or perhaps excruciating) detail in Jackendoff 1983, 1990, 1992a, and 2002a, among many other places. Nevertheless, I have to acknowledge that this is not the leading school of thought in contemporary semantic theory, so it may not prove to be a popular vehicle for studying social cognition.

If, however, one mark of a theory's success is its ability to rigorously address a broad range of questions it was not explicitly designed for and to make connections with seemingly distant disciplines, then the present study vindicates the methodology and philosophy behind Conceptual Semantics. The theory has provided analyses of numerous fundamental concepts of social cognition and theory of mind, and it has shown how they interact with each other in inference and in guiding action. It has related these results to disciplines as disparate as cognitive, clinical, and evolutionary psychology, economics, legal theory, and moral philosophy. It has also led to investigating the relation between language and consciousness (chapter 3). To be sure, these connections have yet to be explored thoroughly, but that is a task that calls for collaboration across disciplinary boundaries, with the expectation of enrichment on both sides. A linguist can't be expected to do it all alone.

At the same time, the analyses developed here have not lost contact with the roots of Conceptual Semantics in explaining how language expresses thought. Through having more precise treatments of the semantics of predicates of perception, evaluation, intention, and value, we have been able to sharpen our understanding of the mapping from semantics to syntax—though not always in ways that would please syntactocentric sensibilities.

To my knowledge, no other extant semantic theory has been able to achieve these sorts of results; nor is any other theory in a position to address the sorts of questions these results bring up. In the end, one's choice of theoretical commitments is always a matter of how to place one's bets. Far more than the arguments for why semantics must be done this way in principle, it is these results that convince me that this is the most productive course for me to pursue in semantic theory.

I am certainly committed to the hypothesis that the formal analysis will find appropriate counterparts in neural activity. But, given the limitations on technology in today's neuroscience, I don't expect to be proven right or wrong within my lifetime. This is neither reason to be discouraged nor reason to disregard what's going on in the rest of cognitive neuroscience.

12.2 Theory of Mind and Social Cognition: What's Innate, and What's Special to Humans?

One of the fundamental motivating issues for linguistic theory is what parts of the adult language capacity are the product of learning and what parts are due to the inherent structure of the brain—including an ability to acquire this particular sort of knowledge. In turn, it is of concern to determine which of the latter parts are special adaptations in humans and, among these, which are specific to the language capacity.

As suggested in chapter 5, similar questions can be asked of social cognition. Here the situation is still more ramified, because other primates have forms of social cognition as well. So the questions may be stated like this:

- Which parts of the adult social/cultural capacity are the product of learning, and which are due to the inherent structure of the brain?
- Of the latter, which parts are shared with other primates, and which are special to humans?
- Independently, which parts are specific to social/cultural cognition, and which are more general cognitive capacities?

This situation creates four possibilities for innate aspects of social cognition:

Type 1: General-purpose in humans and other primates
Type 2: General-purpose in humans, but not present in other primates
Type 3: Special-purpose for social cognition in humans and other primates
Type 4: Special-purpose for social cognition, present in humans only

These divisions are always a matter of degree, in that a special capacity in humans is often the result of "tuning" or amplifying a capacity present already in primates, and a special capacity in any event usually draws on more general-purpose components. For instance, the special properties of the human hand are basically a consequence of changing a few proportions in the shared anatomy of the primate hand, and there is nothing special about the constitution of the bones, muscles, and connective tissue per se that makes them specific to hands. With these issues in mind, let us review our investigation of theory of mind and social cognition.

Theory of mind came up on several occasions in part II: in chapter 6, the distinction between looking (no theory of mind) and seeing (theory of mind); and in chapters 7 and 9, the distinction between objective evaluations and values on one hand (no theory of mind) and subjective evaluations and values on the other hand (theory of mind). Chapter 8 introduced theory of mind in the understanding of intention (including the intentional stance), of commitment to norms, and of belief. In each case, the theory-of-mind predicates involve a conceptualization (or "conceptual simulation") of some combination of valuation features in consciousness, in the sense developed in chapter 3. Without such a conceptualization, an organism can have experiences with the requisite valuation, but it cannot *think about* these aspects of its experience. With this conceptualization in place, an organism can not only reason about its own experience, but also reason about the experiences of others—without experiencing them. This is precisely what theory of mind is supposed to allow one to do, and it emerges from the proposed formalization and its relation to the proposed theory of consciousness.

What's innate here? These conceptualizations are specialized bits within the repertoire of conceptual structure. In principle, an organism could have one of them and not another, so theory of mind is potentially a piecemeal affair. One might wonder to what extent they are dependent on having language. Certainly, language gives a learner more reliable access to these concepts, but an organism lacking the ability to conceptualize valuation features would be incapable of learning theory-of-mind vocabulary such as *see* and *intend*, and would be confined to understanding and predicting perceptible behavior. Such an organism would also be incapable of learning linguistic constructions that involve interpersonal differences in evaluation, such as the difference between *interesting to you* and *interesting to me*; things would have to be just *interesting* or not. I take it to be an empirical question whether nonlinguistic species can use such concepts. That is, it is not clear whether theory-of-mind predicates are of type 3 or type 4 (or whether there are some of each).

Turning to social cognition: The whole possibility of social cognition is predicated on the existence of a distinct tier in conceptual structure in which persons interact and over which social inferences are defined. This tier is linked to the physical plane of conceptual structure through general mechanisms common to the linking of all sorts of planes in cognition, but the tier itself and its properties are specific to social cognition. It is likely that other social animals have some counterpart of this tier in their conceptual structure, through which they treat conspecifics as having a special status and engage in specialized sorts of interactions. Nevertheless, the human social tier is undoubtedly far more elaborate. In particular, I have argued that the personal plane in humans is responsible for such peculiar but culturally ubiquitous phenomena as the conceptualization of supernatural beings and belief in life after death and/or reincarnation.

One of the recurring factors throughout our discussion of social cognition has been group membership and its logic: one treats members of one's own group more favorably than members of other groups. Humans must learn which groups they belong to, who else is in these groups, what other groups they interact with and who belongs to them, and what the customs are for interaction within groups and between groups (including regulation of intragroup variation and intergroup tolerance). This is indeed a great deal to learn. But the basic logic of groups that underlies all this learning is likely innate. This too is an aspect of social cognition that is widespread among animals but hugely elaborated in humans, where it includes the possibility of embedded and overlapping groups and the conception of a group as a superindividual. We have reached similar conclusions with respect to other relationships such as kinship and dominance: all of these are some combination of type 3 factors overlaid with type 4 elaborations.

Social cognition is of course deeply involved in any sort of cooperative activity. Cooperation of the human sort depends on the concept of a joint task. In a joint task, the intention is not just "I will do such-and-such," but "*We* will do such-and-such, and our respective roles in the task are such-and-such." Joint activity requires initiating steps of offer and uptake, which establish a joint intention, a simulated "sharing of minds" that brings with it mutual obligations to carry out the task. The execution of a joint task presents not only problems of coordination, but also opportunities for defection, hence requiring extensive use of theory of mind and "cheater detection." It is hard to see joint intention as other than a cognitive specialization involved in social cognition, as it is strictly concerned with interactions with other individuals. Do animals have such a concept? The evidence is mixed, but in any event humans make far more extensive use of it.

Consider next the systems of value. Affective (A-)value, how good or bad something feels, is not a social notion. Animals may well have the objective form of A-value in their conceptual repertoire, as it is involved in choosing among different possible courses of action. They probably lack the subjective version, though this awaits investigation. On the other hand, the explicit concept of A-value is probably laid on top of more primitive systems that guide action in organisms lacking a developed conceptual system.

Similar considerations pertain to quality (Q-value). Insofar as an organism strives for the best of anything—the best food, the best habitat, the best mate—it is operating with a notion of Q-value. Again, it is likely that an explicit concept of Q-value is built on top of more primitive systems, since selective behavior is hardly confined to organisms with fancy cognitive systems.

Utility (U-value) is also not inherently a social notion. Nevertheless, we humans learn a great deal about U-values from our culture: all the cultural and technological apparatus for how to provide for our welfare. Unlike A-value, which may be widespread among animals, I find it less plausible that U-value is an explicit concept in any but a few species. My impression is that in the situations where animals need to take into account the future benefits of an action, evolution has made the action feel good to them, so the operative value turns out to be A-value: the quintessential cases are eating and sex. To establish that an animal is genuinely operating in terms of U-value, I think it would be necessary to show that it defers pleasure in favor of greater anticipated pleasure in the future. Is a bird or a chimpanzee who is building a nest doing it with a sense of its utility, or does it just feel good to do so? The answers might turn out differently depending on species, of course.

To the extent that an animal guards resources (e.g. territory) or gathers and hoards food, we may be able to say that it has a notion of resource (R-)value. On the other hand, the virtue of the concept of R-value in humans is the fungibility of different sorts of valuable objects: the ability to compare their values and exchange them for each other. Without the notion of agreed-upon exchange of objects, as far as I can tell completely absent in other species, the concept of R-value is quite thin. Thus I would be inclined to consider R-value a type 4 concept, though not without type 3 precursors. There is of course a great deal for a human to learn about R-values of particular objects, especially those that are fraught with social meaning. Still, like the other sorts of value, one could not learn these things without a conceptual framework that included the possibility of R-values and their characteristic inference patterns.

There are three types of value that pertain to other persons: prowess (P-value), esteem (E-value), and personal normative (PN-)value. The element in animal societies that appears to undergird these three is dominance. Prowess is the most closely related, but unlike sheer dominance, it is measured relative to a particular activity: one can be good at swimming or linguistics without achieving overall dominance. Esteem is a composite of dominance and a variety of other factors, including group membership and PN-value. It is not clear to me whether primate societies involve such differentiated elaborations of dominance. Displays of esteem and disrespect, highly developed in human cultures and demanding a great deal of learning and often constant vigilance, are elaborations of primate displays of dominance and submission.

PN-value of course is derived from the normative (N-)value of an individual's actions. In turn, N-value is highly differentiated into moral/ethical value, religious value, manners/etiquette, and perhaps others. Again, an individual must learn a vast amount of information about what actions the culture considers of N-value (positive or negative), as well as what constitute acceptable actions to take in response to others' N-valued actions. These normative prescriptions may be focused around certain prominent and ubiquitous issues such as not hurting people, showing respect, conformity to group standards of appearance and behavior, and differential treatment of kin versus nonkin and group members versus nonmembers. But there is a huge amount of variation in how these prescriptions are realized—not unlike the situation in language variation. I don't see a precursor of N-value in primate societies, but I'm willing to be persuaded that one can be found through careful observation and experiment.

What makes N-value particularly complex is its multiple roles in inference. What has often been seen at its evolutionary root is the Machiavellian desire to increase the esteem in which one is held, which in turn increases opportunities for profitable interaction with others. But when N-values are internalized, it can just feel good to conform to them (they carry along with them A-value); or one may conform to norms "because it's the right thing to do"—they define one's duty regardless of inherent unpleasantness. Paradoxically, if you are discovered to be doing good things for the sake of enhancing your reputation, people think less of you, not more. The consequence is cycles of deception and perhaps self-deception, and more need for cheater detection. Here there is even less antecedent in the primates.

Values in turn are crucial to reciprocal action of all sorts: freely chosen reciprocation, retaliation, and restitution; reciprocal, retaliatory, and restitutive displays of (dis)respect for P- and PN-value; and agreed-upon exchanges. In each case, the paired actions are compared along a scale of value, and an N-value is placed on conducting the exchange in proper fashion. In the case of reciprocation and restitution, there is a standard positive N-value for reciprocating; the N-value of retaliation varies from circumstance to circumstance and from culture to culture. In the case of agreed-upon exchanges, there are instead mutual obligations that affect the utility of defaulting on the exchange.

These frames are culturally ubiquitous, though there are lots of cross-cultural differences in how they are applied. There are less differentiated and less flexible antecedents in other species' use of reciprocal altruism, retaliation, and restitution (e.g. reconciliation). In particular, what has often been called "cooperative reciprocal altruism" in the literature is actually freely chosen reciprocation, which does not require cooperation at all. What is remarkable in humans is the broad generality of all these sorts of actions and the high degree of differentiation among them.

Implicit or explicit rules governing the assignment of N-value are distinct from a further sort of normative principle discussed here: rights and obligations. Rights and obligations are conceptualized as objective social constraints on action. They are remarkably abstract, in that they derive their force only from what happens when they are violated: other rights come into existence, which in turn are affordances for retaliatory action. In turn, ownership, joint action, contracts, and legal systems all depend on a grasp of rights and obligations. The social notion of authority, another elaboration of dominance, governs much of the distribution of rights and obligations. And theories of government are based on how authority is distributed and justified. Here the complexity is such that it is hard to see nonhuman antecedents.

To sum up, it is not as though there are all-or-nothing modules for theory of mind and social cognition that either are or are not specific to humans. Rather, there is a repertoire of abstract concepts that invoke the social plane, some of which are specific to humans and some of which are more highly differentiated in humans. They hang together as a functional module through their inferential interactions on the social plane, and each gets a certain amount of its content through its role in these interactions.

The present study has only begun to scratch the surface. Many questions can be raised about details of the analysis; many connections to other enterprises have only been hinted at; and many aspects of social cognition, for instance emotion, remain untouched. My speculations about innateness and species-specificity are at the moment just that. One might justifiably wonder whether some of the social concepts and inferences explored here are necessarily innate, as I believe, or whether they might instead be general-purpose cognition's optimal solutions to inherent problems of coordinating large groups—as it were, nothing but successful “memes.” One might also wonder how essential language is to the richness of the system; surely language and social cognition are deeply intertwined.

At the same time, I believe it has been useful to attempt a fairly synoptic view of the concepts involved in social cognition. Much of the recent work on social cognition within cognitive science has focused on such areas as affect recognition, false belief tasks, rational choice, heuristics, cheater detection, morality, and religion, without much analysis of the larger cognitive context in which these phenomena reside. Here I have tried to proceed from the big picture inward; the connections to these more specialized domains have fallen out here and there, more or less casually. I do not mean thereby to minimize the importance of these domains. Rather, I hope that the connections built here are fruitful in attempting to unify the inquiry, and that researchers in these domains—along with many other domains—will come to consider the whole enterprise a joint task.

References

Adams, Fred. 2003. Semantic paralysis. (Commentary on Jackendoff 2003.) *Behavioral and Brain Sciences* 26, 666–667.

Ainslie, George. 2001. *Breakdown of will*. Cambridge: Cambridge University Press.

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17, 673–711.

Akerlof, George A., and Janet L. Yellen. 1993. The fair wage–effort hypothesis in unemployment. In Hechter, Nadel, and Michod 1993, 107–134.

Akmajian, Adrian. 1984. Sentence types and the form-function fit. *Natural Language and Linguistic Theory* 2, 1–23.

Alexander, Richard D. 1987. *The biology of moral systems*. New York: Aldine de Gruyter.

Allen, Brooke. 2005. Our godless constitution. *The Nation*, February 21, 2005, 14–20.

Anderson, Stephen. 1977. Comments on the paper by Wasow. In Culicover, Wasow, and Akmajian 1977, 361–377.

Aronoff, Mark. 1980. Contextuals. *Language* 56, 744–758.

Atran, Scott. 2004. *In gods we trust: The evolutionary landscape of religion*. Oxford: Oxford University Press.

Atran, Scott, Douglas L. Medin, and Norbert O. Ross. 2005. The cultural mind: Environmental decision making and cultural modeling within and across populations. *Psychological Review* 112, 744–776.

Baars, Bernard J. 1988. *A cognitive theory of consciousness*. New York: Cambridge University Press.

Baars, Bernard J. 1997. Understanding subjectivity: Global Workspace Theory and the resurrection of the self. In Shear 1997, 241–248.

Baars, Bernard J. 2003. Working memory requires conscious processes, not vice versa: A Global Workspace account. In Osaka 2003, 11–26.

Bach, Kent, and Robert Harnish. 1979. *Linguistic communication and speech acts*. Cambridge, Mass.: MIT Press.

Baddeley, Alan. 1986. *Working memory*. Oxford: Clarendon Press.

Badler, Norman I., Rama Bindiganavale, Jan Allbeck, William Schuler, Liwei Zhao, and Martha Palmer. 2000. A parameterized action representation for virtual human agents. In Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, eds., *Embodied conversational agents*, 256–284. Cambridge, Mass.: MIT Press.

Baillargeon, Renée. 1986. Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition* 23, 21–41.

Baker, C. L., and John J. McCarthy, eds. 1981. *The logical problem of language acquisition*. Cambridge, Mass.: MIT Press.

Baker, Mark. 1988. *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.

Bangerter, Adrian, and Herbert H. Clark. 2003. Navigation joint projects with dialogue. *Cognitive Science* 27, 195–225.

Barth, Fredrik. 1993. Are values real? The enigma of naturalism in the anthropological imputation of values. In Hechter, Nadel, and Michod 1993, 31–46.

Barwise, Jon, and John Perry. 1983. *Situations and attitudes*. Cambridge, Mass.: MIT Press.

Barzun, Jacques. 1958. *Darwin, Marx, Wagner: Critique of a heritage*. 2nd ed. New York: Doubleday Anchor Books.

Beckman, Mary, and Janet Pierrehumbert. 1986. Intonational structure in English and Japanese. *Phonology* 3, 255–309.

Belletti, Adriana, and Luigi Rizzi. 1988. Psych-verbs and θ-theory. *Natural Language and Linguistic Theory* 6, 291–352.

Bellugi, Ursula, Howard Poizner, and Edward S. Klima. 1989. Language, modality, and the brain. *Trends in Neurosciences* 12, 380–388.

Bellugi, Ursula, Paul P. Wang, and Terry L. Jernigan. 1994. Williams syndrome: An unusual neuropsychological profile. In Sarah H. Broman and Jordan Grafman, eds., *Atypical cognitive deficits in developmental disorders: Implications for brain function*, 23–56. Hillsdale, N.J.: Erlbaum.

Berger, Peter L., and Thomas Luckmann. 1966. *The social construction of reality*. Garden City, N.Y.: Doubleday.

Bickerton, Derek. 1981. *Roots of language*. Ann Arbor, Mich.: Karoma.

Biederman, Irving. 1987. Recognition-by-components: A theory of human image understanding. *Psychological Review* 94, 115–147.

Bierwisch, Manfred. 1967. Some semantic universals of German adjectivals. *Foundations of Language* 3, 1–36.

Bierwisch, Manfred. 1969. On certain problems of semantic representation. *Foundations of Language* 5, 153–184.

Bindiganavale, Rama, William Schuler, Jan M. Allbeck, Norman I. Badler, Aravind K. Joshi, and Martha Palmer. 2000. Dynamically altering agent behaviors using natural language instructions. In *Proceedings of the 4th International Conference on Autonomous Agents*, 293–300. New York: ACM Press.

Block, Ned. 1995. On a confusion about the function of consciousness. *Behavioral and Brain Sciences* 18, 227–287.

Bloom, Paul. 2000. *How children learn the meanings of words*. Cambridge, Mass.: MIT Press.

Bloom, Paul. 2004. *Descartes' baby: How the science of child development explains what makes us human*. New York: Basic Books.

Bloom, Paul, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, eds. 1996. *Language and space*. Cambridge, Mass.: MIT Press.

Bloom, Paul, and Csaba Veres. 1999. The perceived intentionality of groups. *Cognition* 71, B1–B9.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.

Bowerman, Melissa. 1996. Learning how to structure space for language: A cross-linguistic perspective. In Bloom et al. 1996, 385–436.

Boyd, Robert, and Peter J. Richerson. 2005. *The origin and evolution of cultures*. Oxford: Oxford University Press. [contains essays originally published between 1989 and 2003]

Boyer, Pascal. 2001. *Religion explained*. New York: Basic Books.

Bratman, Michael E. 1987. *Intentions, plans, and practical reason*. Cambridge, Mass.: Harvard University Press.

Bratman, Michael E. 1990. What is intention? In Cohen, Morgan, and Pollack 1990, 15–31.

Bratman, Michael E. 1999. *Faces of intention*. Cambridge: Cambridge University Press. [contains essays previously published between 1985 and 1998]

Bresnan, Joan W., ed. 1982. *The mental representation of grammatical relations*. Cambridge, Mass.: MIT Press.

Bresnan, Joan W. 2001. *Lexical-functional syntax*. Oxford: Blackwell.

Bresnan, Joan W., and Jonni Kanerva. 1989. Locative inversion in Chicheŵa: A case study of factorization in grammar. *Linguistic Inquiry* 20, 1–50.

Brown, Donald. 1991. *Human universals.* New York: McGraw-Hill.

Bruner, Jerome. 1983. *In search of mind.* New York: Harper and Row.

Bybee, Joan, and James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22, 381–410.

Byrne, Richard W., and Andrew Whiten, eds. 1988. *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Clarendon Press.

Carey, Susan. 1985. *Conceptual change in childhood.* Cambridge, Mass.: MIT Press.

Carnap, Rudolf. 1939. *Foundations of logic and mathematics*. Chicago: University of Chicago Press.

Carter, Richard. 1976. Some linking regularities. In *On linking: Papers by Richard Carter*, ed. by Beth Levin and Carol Tenny. Cambridge, Mass.: MIT, Center for Cognitive Science Lexicon Project.

Castañeda, Hector-Neri. 1975. *Thinking and doing*. Dordrecht: Reidel.

Cavalli-Sforza, Luigi L. 2001. *Genes, peoples, and languages*. Berkeley and Los Angeles: University of California Press.

Cavanagh, Patrick, and George A. Alvarez. 2005. Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences* 9, 349–354.

Cavanagh, Patrick, Angela T. Labianca, and Ian M. Thornton. 2001. Attention-based visual routines: Sprites. *Cognition* 80, 47–60.

Chalmers, David. 1997. Facing up to the problem of consciousness. In Shear 1997, 9–30.

Cheney, Dorothy, and Robert Seyfarth. 1990. *How monkeys see the world*. Chicago: University of Chicago Press.

Chierchia, Gennaro, and Sally McConnell-Ginet. 1990. *Meaning and grammar: An introduction to semantics*. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1966. *Cartesian linguistics*. New York: Harper and Row.

Chomsky, Noam. 1972. *Studies on semantics in generative grammar*. The Hague: Mouton.

Chomsky, Noam. 1973. Constraints on transformations. In Stephen Anderson and Paul Kiparsky, eds., *A festschrift for Morris Halle*, 232–286. New York: Holt, Rinehart and Winston.

Chomsky, Noam. 1977. On *wh*-movement. In Culicover, Wasow, and Akmajian 1977, 71–132.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 2002. *On nature and language*. Cambridge: Cambridge University Press.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36, 1–22.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Chomsky, Noam, and Howard Lasnik. 1993. The theory of principles and parameters. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Vennemann, eds., *Syntax: An international handbook of contemporary research*, 506–569. Berlin: de Gruyter. Also in Chomsky 1995, 13–127.

Churchland, Paul M. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78, 67–90.

Churchland, Paul M., and Patricia S. Churchland. 2003. Recent work on consciousness: Philosophical, theoretical, and empirical. In Osaka 2003, 123–138.

Clahsen, Harald, and Mayella Almazan. 1998. Syntax and morphology in Williams syndrome. *Cognition* 68, 167–198.

Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.

Clark, Thomas W. 1997. Function and phenomenology: Closing the explanatory gap. In Shear 1997, 45–60.

Cohen, Philip R., and Hector J. Levesque. 1991. Teamwork. *Noûs* 25, 487–512.

Cohen, Philip R., Jerry Morgan, and Martha E. Pollack, eds. 1990. *Intentions in communication*. Cambridge, Mass.: MIT Press.

Collins, A. M., and M. R. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8, 240–247.

Conison, Jay. 1997. The pragmatics of *promise*. *Canadian Journal of Law and Jurisprudence* 10, 273–322.

Corballis, Michael C. 1991. *The lopsided ape*. Oxford: Oxford University Press.

Cosmides, Leda. 1989. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition* 31, 187–276.

Cosmides, Leda, and John Tooby. 1992. Cognitive adaptations for social exchange. In Jerome Barkow, Leda Cosmides, and John Tooby, eds., *The adapted mind*, 163–228. New York: Oxford University Press.

Coventry, Kenny R., and Simon C. Garrod. 2004. *Saying, seeing, and acting: The psychological semantics of spatial prepositions*. New York: Psychology Press.

Crick, Francis. 1994. *The astonishing hypothesis: The scientific search for the soul*. New York: Charles Scribner's Sons.

Crick, Francis, and Cristof Koch. 1990. Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences* 2, 263–275.

Crick, Francis, and Cristof Koch. 1995. Are we aware of neural activity in primary visual cortex? *Nature* 375, 121–123.

Culicover, Peter. 1972. OM-sentences. *Foundations of Language* 8, 199–236.

Culicover, Peter. 1999. *Syntactic nuts: Hard cases in syntax*. Oxford: Oxford University Press.

Culicover, Peter, and Ray Jackendoff. 2005. *Simpler syntax*. Oxford: Oxford University Press.

Culicover, Peter, Thomas Wasow, and Adrian Akmajian, eds. 1977. *Formal syntax*. New York: Academic Press.

Cummins, Denise D. 1996a. Evidence for the innateness of deontic reasoning. *Mind and Language* 11, 160–190.

Cummins, Denise D. 1996b. Evidence of deontic reasoning in 3- and 4-year-old children. *Memory and Cognition* 24, 823–829.

Curtiss, Susan. 1977. *Genie: A linguistic study of a modern-day "wild child."* New York: Academic Press.

Cutler, Anne, and Charles Clifton, Jr. 1999. Comprehending spoken language: A blueprint of the listener. In Colin M. Brown and Peter Hagoort, eds., *The neurocognition of language*, 123–166. Oxford: Oxford University Press.

Cutting, James. 1981. Six tenets for event perception. *Cognition* 10, 71–78.

Damasio, Antonio R. 1994. *Descartes' error: Emotion, reason, and the human brain*. New York: G. P. Putnam's Sons.

Damasio, Antonio R. 2000. A neurobiology for consciousness. In Metzinger 2000, 111–120.

Dawkins, Richard. 1989. *The selfish gene*. New ed. Oxford: Oxford University Press.

Deacon, Terrence W. 1997. *The symbolic species*. New York: W. W. Norton.

Decety, Jean, and Jessica A. Sommerville. 2003. Shared representations between self and other: A social cognitive neuroscience view. *Trends in Cognitive Sciences* 7, 527–533.

Degler, Carl N. 1991. *In search of human nature: The decline and revival of Darwinism in American social thought*. New York: Oxford University Press.

DeGraff, Michel, ed. 1999. *Language creation and language change: Creolization, diachrony, and development*. Cambridge, Mass.: MIT Press.

Dehaene, Stanislas. 1997. *The number sense: How the mind creates mathematics.* Oxford: Oxford University Press.

Dehaene, Stanislas, ed. 2001. The cognitive neuroscience of consciousness. Special issue, *Cognition* 79: 1–2.

Dehaene, Stanislas, Jean-Pierre Changeux, Lionel Naccache, Jérôme Sackur, and Clair Sergent. 2006. Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences* 10, 204–211.

Dehaene, Stanislas, and Lionel Naccache. 2001. Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. In Dehaene 2001, 1–37.

Dell, Gary S., Lisa K. Burger, and William R. Svec. 1997. Language production and serial order: A functional analysis and a model. *Psychological Review* 104, 123–147.

Dennett, Daniel C. 1984. *Elbow room: The varieties of free will worth wanting.* Cambridge, Mass.: MIT Press.

Dennett, Daniel C. 1987. *The intentional stance*. Cambridge, Mass.: MIT Press.

Dennett, Daniel C. 1991. *Explaining consciousness.* New York: Little, Brown.

Dennett, Daniel C. 2001. Are we explaining consciousness yet? In Dehaene 2001, 221–237.

Dennett, Daniel C. 2003. *Freedom evolves*. New York: Viking.

de Waal, Frans B. M. 1996. *Good natured: The origins of right and wrong in humans and other animals*. Cambridge, Mass.: Harvard University Press.

Diamond, Jared. 1997. *Guns, germs, and steel: The fates of human societies.* New York: W. W. Norton.

Donald, Merlin. 1998. Preconditions for the evolution of protolanguages. In Michael C. Corballis and Stephen E. G. Lea, eds., *The descent of mind*, 355–365. Oxford: Oxford University Press.

Doris, John M., and Stephen P. Stich. 2005. As a matter of fact: Empirical perspectives on ethics. In Frank Jackson and Michael Smith, eds., *The Oxford handbook of contemporary philosophy*, 114–152. Oxford: Oxford University Press.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67, 547–619.

Driver, Jon, Greg Davis, Charlotte Russell, Massimo Turatto, and Elliot Freeman. 2001. Segmentation, attention and phenomenal visual objects. *Cognition* 80, 61–95.

Driver, Jon, and Patrik Vuilleumier. 2001. Perceptual awareness and its loss in unilateral neglect and extinction. In Dehaene 2001, 39–88.

Dufva, Hannele, and Mika Lähteenmäki. 1996. But who killed Harry? A dialogical approach to language and consciousness. *Pragmatics and Cognition* 4, 35–53.

Edelman, Gerald M., and Giulio Tononi. 2000. Reentry and the dynamic core: Neural correlates of conscious experience. In Metzinger 2000, 139–151.

Ehrenreich, Barbara, and Janet McIntosh. 1997. The new creationism: Biology under attack. *The Nation*, June 9, 1997, 11–16.

Eibl-Eibesfeldt, Irenäus. 1989. *Human ethology*. New York: Aldine de Gruyter.

Ekman, Paul, and Richard J. Davidson, eds. 1994. *The nature of emotion*. New York: Oxford University Press.

Ellickson, Robert C. 1991. *Order without law: How neighbors settle disputes*. Cambridge, Mass.: Harvard University Press.

Elman, Jeffrey. 1990. Finding structure in time. *Cognitive Science* 14, 179–211.

Elman, Jeffrey, Elizabeth Bates, Mark Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. Cambridge, Mass.: MIT Press.

Fauconnier, Gilles. 1985. *Mental spaces: Aspects of meaning construction in natural language*. Cambridge, Mass.: MIT Press.

Fehr, Ernst, and Urs Fischbacher. 2004. Social norms and human cooperation. *Trends in Cognitive Sciences* 8, 185–190.

ffytche, Dominic. 2002. Neural codes for conscious vision. *Trends in Cognitive Sciences* 6, 493–495.

Fikes, Richard, and Nils J. Nilsson. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2, 189–208.

Fillmore, Charles. 1965. *Indirect object constructions and the ordering of transformations*. The Hague: Mouton.

Fillmore, Charles. 1988. The mechanisms of "Construction Grammar." In Shelley Axmaker, Annie Jaisser, and Helen Singmaster, eds., *Proceedings of the 14th Annual Meeting of the Berkeley Linguistics Society*, 35–55. Berkeley, Calif.: University of California, Berkeley Linguistics Society.

Fillmore, Charles, and Beryl Atkins. 1992. Toward a frame-based lexicon: The semantics of RISK and its neighbors. In Adrienne Lehrer and Eva Kittay, eds., *Frames, fields, and contrasts*, 75–102. Hillsdale, N.J.: Erlbaum.

Fillmore, Charles, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64, 501–538.

Fischer, Susan D., and Patricia Siple, eds. 1990. *Theoretical issues in sign language research 1*. Chicago: University of Chicago Press.

Fiske, Alan Page. 1991. *Structures of social life: The four elementary forms of human relations*. New York: Free Press.

Fitch, W. Tecumseh. 2000. The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4, 258–267.

Fitch, W. Tecumseh, Marc D. Hauser, and Noam Chomsky. 2005. The evolution of the language faculty: Clarifications and implications (Reply to Pinker and Jackendoff). *Cognition* 97, 179–210.

Flanagan, J. Randall, and Roland S. Johansson. 2003. Action plans used in action observation. *Nature* 424, 769–771.

Flohr, Hans. 2000. NMDA receptor-mediated computational processes and phenomenal consciousness. In Metzinger 2000, 245–258.

Flynn, Suzanne, and Wayne O'Neil, eds. 1988. *Linguistic theory in second language acquisition*. Dordrecht: Reidel.

Fodor, Jerry A. 1970. Three reasons for not deriving "kill" from "cause to die." *Linguistic Inquiry* 1, 429–438.

Fodor, Jerry A. 1975. *The language of thought*. Cambridge, Mass.: Harvard University Press.

Fodor, Jerry A. 1983. *The modularity of mind*. Cambridge, Mass.: MIT Press.

Fodor, Jerry A. 1987. *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, Mass.: MIT Press.

Fodor, Jerry A. 1998. *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Fodor, Jerry A., Thomas Bever, and Merrill Garrett. 1974. *The psychology of language*. New York: McGraw-Hill.

Fodor, Jerry A., and Ernest Lepore. 1992. *Holism*. Oxford: Blackwell.

Forrester, James W. 1996. *Being good and being logical: Philosophical groundwork for a new deontic logic*. London: M. E. Sharpe.

Frank, Robert, and Anthony Kroch. 1995. Generalized transformations and the theory of grammar. *Studia Linguistica* 49, 103–151.

Frege, Gottlob. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100, 25–50. English translation in Peter Geach and Max Black, eds., *Translations from the philosophical writings of Gottlob Frege*, 56–78. Oxford: Blackwell, 1952.

Gallese, Vittorio, Christian Keysers, and Giacomo Rizzolatti. 2004. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences* 8, 396–403.

Gardner, Howard. 1983. *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Geertz, Clifford. 1973. *The interpretation of cultures*. New York: Basic Books.

Gergely, G., Z. Nádasdy, G. Csibra, and S. Bíró. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56, 165–193.

Ghomeshi, Jila, Ray Jackendoff, Nicole Rosen, and Kevin Russell. 2004. Contrastive focus reduplication in English (The salad-salad paper). *Natural Language and Linguistic Theory* 22, 307–357.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group. 2000. *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gilbert, Margaret. 1989. *On social facts*. Princeton, N.J.: Princeton University Press.

Ginzburg, Jonathan, and Ivan A. Sag. 2000. *Interrogative investigations: The form, meanings, and use of English interrogatives*. Stanford, Calif.: CSLI Publications.

Gleitman, Lila R., Henry Gleitman, Carol Miller, and Ruth Ostrin. 1996. Similar, and similar concepts. *Cognition* 58, 321–376.

Goffman, Erving. 1974. *Frame analysis: An essay on the organization of experience*. Cambridge, Mass.: Harvard University Press.

Goldberg, Adele. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele. 2006. *Constructions at work*. Oxford: Oxford University Press.

Goldman, Alvin. 1993. Ethics and cognitive science. *Ethics* 103, 337–360.

Goldsmith, John. 1979. *Autosegmental phonology*. New York: Garland Press.

Goldsmith, John, ed. 1995. *Handbook of theoretical phonology*. Oxford: Blackwell.

Goodall, Jane van Lawick. 1971. *In the shadow of man*. New York: Dell.

Goodenough, Ward H. 1970. *Description and comparison in cultural anthropology*. Chicago: Aldine.

Gopnik, Myrna. 1999. Some evidence for impaired grammars. In Ray Jackendoff, Paul Bloom, and Karen Wynn, eds., *Language, logic, and concepts*, 263–283. Cambridge, Mass.: MIT Press.

Gray, Charles M., Peter König, Andreas K. Engel, and Wolf Singer. 1989. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337.

Greene, Joshua. 2003. From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? *Nature Reviews/Neuroscience* 4, 847–850.

Grimshaw, Jane. 1990. *Argument structure*. Cambridge, Mass.: MIT Press.

Gross, Steven. 2005. The nature of semantics: On Jackendoff's arguments. *The Linguistic Review* 22, 249–270.

Grosz, Barbara J., and Candace L. Sidner. 1999. Plans for discourse. In Cohen, Morgan, and Pollack 1999, 417–444.

Gruber, Jeffrey. 1965. Studies in lexical relations. Doctoral dissertation, MIT. Published as part of *Lexical structures in syntax and semantics*. Amsterdam: North-Holland, 1976.

Halle, Morris, and William Idsardi. 1995. Stress and metrical structure. In Goldsmith 1995, 403–443.

Halle, Morris, and Alec Marantz. 1993. Distributed Morphology. In Kenneth Hale and Samuel Jay Keyser, eds., *The view from Building 20*, 111–176. Cambridge, Mass.: MIT Press.

Hameroff, Stuart, and Roger Penrose. 1997. Conscious events as orchestrated space-time selections. In Shear 1997, 177–195.

Hamilton, William D. 1964. The genetical evolution of social behaviour (I and II). *Journal of Theoretical Biology* 7, 1–16; 17–52.

Hardcastle, Valerie Gray. 2003. Attention versus consciousness: A distinction with a difference. In Osaka 2003, 105–120.

Harman, Gilbert. 2000. *Explaining value*. Oxford: Oxford University Press.

Harris, Paul, and María Núñez. 1997. Children's understanding of permission and obligation. In Leslie Smith, Julie Dockrell, and Peter Tomlinson, eds., *Piaget, Vygotsky and beyond*, 211–223. London: Routledge.

Harris, Randy Allen. 1993. *The linguistics wars*. New York: Oxford University Press.

Hauser, Marc D. 2000. *Wild minds: What animals really think*. New York: Henry Holt.

Hauser, Marc D. 2006. *Moral minds: How nature designed our sense of right and wrong*. New York: HarperCollins/Ecco Press.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579.

Hechter, Michael, Lynn Nadel, and Richard E. Michod, eds. 1993. *The origin of values*. New York: Aldine de Gruyter.

Heider, Fritz, and Marianne Simmel. 1944. An experimental study of apparent behavior. *American Journal of Psychology* 57, 243–249.

Heim, Irene, and Angelika Kratzer. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Heny, Frank, ed. 1981. *Ambiguities in intensional contexts*. Dordrecht: Reidel.

Herrnstein, Richard J. 1993. Behavior, reinforcement, and utility. In Hechter, Nadel, and Michod 1993, 137–152.

Herrnstein, Richard J., and Charles Murray. 1994. *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

Herskovits, Annette. 1986. *Language and spatial cognition*. Cambridge: Cambridge University Press.

Higginbotham, James. 2003. Jackendoff's conceptualism. (Commentary on Jackendoff 2003.) *Behavioral and Brain Sciences* 26, 680–681.

Hirschfeld, Lawrence. 1996. *Race in the making: Cognition, culture, and the child's construction of human kinds*. Cambridge, Mass.: MIT Press.

Hobbs, Jerry. 1999. Artificial intelligence and collective intentionality: Comments on Searle and on Grosz and Sidner. In Cohen, Morgan, and Pollack 1999, 445–459.

Hockett, Charles F. 1960. The origin of speech. *Scientific American* 203, 88–111.

Hoffman, Martin L. 1987. The contribution of empathy to justice and moral judgment. In Nancy Eisenberg and Janet Strayer, eds., *Empathy and its development*, 47–80. Cambridge: Cambridge University Press.

Hofstadter, Douglas. 1979. *Gödel, Escher, Bach*. New York: Basic Books.

Horgan, Terence, and John Tienson. 2006. Cognition needs syntax but not rules. In Robert J. Stainton, ed., *Contemporary debates in cognitive science*, 147–158. Malden, Mass.: Blackwell.

Horn, Laurence. 1993. Economy and redundancy in a dualistic model of natural language. In Susanna Shore and Maria Vilkuna, eds., *SKY 1993*, *Yearbook of the Linguistic Association of Finland*, 31–72. Helsinki: University of Helsinki, Department of General Linguistics.

Huck, Geoffrey, and John Goldsmith. 1995. *Ideology and linguistic theory*. Chicago: University of Chicago Press.

Humphries, Glyn W., Emer M. E. Forde, and M. Jane Riddoch. 2001. The planning and execution of everyday actions. In Brenda Rapp, ed., *The handbook of cognitive neuropsychology: What deficits reveal about the human mind*, 565–589. Philadelphia: Psychology Press.

Hut, Piet, and Roger Shepard. 1997. Turning 'the Hard Problem' upside down and sideways. In Shear 1997, 305–322.

Iwasaki, Syoichi. 1993. Spatial attention and two modes of visual consciousness. *Cognition* 49, 211–233.

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1976. Toward an explanatory semantic representation. *Linguistic Inquiry* 7, 89–150.

Jackendoff, Ray. 1983. *Semantics and cognition.* Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1985. Believing and intending: Two sides of the same coin. *Linguistic Inquiry* 16, 445–459.

Jackendoff, Ray. 1987. *Consciousness and the computational mind.* Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1990. *Semantic structures*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1991. Parts and boundaries. *Cognition* 41, 9–45.

Jackendoff, Ray. 1992a. *Languages of the mind.* Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1992b. Mme. Tussaud meets the binding theory. *Natural Language and Linguistic Theory* 10, 1–31.

Jackendoff, Ray. 1994. *Patterns in the mind.* New York: Basic Books.

Jackendoff, Ray. 1995. The conceptual structure of intending and volitional action. In Héctor Campos and Paula Kempchinsky, eds., *Evolution and revolution in linguistic theory: Studies in honor of Carlos P. Otero*, 198–227. Washington, D.C.: Georgetown University Press.

Jackendoff, Ray. 1996a. The architecture of the linguistic-spatial interface. In Bloom et al. 1996, 1–30.

Jackendoff, Ray. 1996b. How language helps us think. *Pragmatics and Cognition* 4, 1–24. Slightly revised version appears as Jackendoff 1997a, chap. 8.

Jackendoff, Ray. 1996c. The proper treatment of measuring out, telicity, and possibly even quantification in English. *Natural Language and Linguistic Theory* 14, 305–354.

Jackendoff, Ray. 1997a. *The architecture of the language faculty*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray. 1997b. Twistin' the night away. *Language* 73, 534–559.

Jackendoff, Ray. 1999. The natural logic of rights and obligations. In Jackendoff, Bloom, and Wynn 1999, 67–95.

Jackendoff, Ray. 2002a. *Foundations of language*. Oxford: Oxford University Press.

Jackendoff, Ray. 2002b. Review of Jerry A. Fodor, *The mind doesn't work that way. Language* 78, 164–170.

Jackendoff, Ray. 2003. Précis of Jackendoff 2002a. *Behavioral and Brain Sciences* 26, 651–665.

Jackendoff, Ray. 2007. A parallel architecture perspective on language processing. *Brain Research*.

Jackendoff, Ray. Forthcoming. The role of linguistics in cognitive science: The state of the art. *The Linguistic Review*.

Jackendoff, Ray, and David Aaron. 1991. Review of George Lakoff and Mark Turner, *More than cool reason*. Language 67, 320–338.

Jackendoff, Ray, Paul Bloom, and Karen Wynn, eds. 1999. *Language, logic, and concepts: Essays in honor of John Macnamara*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray, and Fred Lerdahl. 2006. The capacity for music: What's special about it? *Cognition* 100, 33–72.

Jackendoff, Ray, and Steven Pinker. 2005. The nature of the language faculty and its implications for the evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition* 97, 211–225.

Jacobs, Jane. 1994. *Systems of survival: A dialogue on the moral foundations of commerce and politics.* New York: Vintage Books.

James, William. 1890. *The principles of psychology*. New York: Dover [reprint], 1950.

Johnson-Laird, Philip. 1983. *Mental models*. Cambridge: Cambridge University Press.

Kager, René. 1995. The metrical theory of word stress. In Goldsmith 1995, 367–402.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kanwisher, Nancy. 2001. Neural events and perceptual awareness. In Dehaene 2001, 89–113.

Katz, Jerrold. 1966. *The philosophy of language*. New York: Harper and Row.

Katz, Jerrold, and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language* 39, 170–210.

Katz, Jerrold, and Paul M. Postal. 1964. *An integrated theory of linguistic descriptions*. Cambridge, Mass.: MIT Press.

Kautz, Henry. 1990. A circumscriptive theory of plan recognition. In Cohen, Morgan, and Pollack 1990, 105–133.

Kegl, Judy, Ann Senghas, and Marie Coppola. 1999. Creations through contact: Sign language emergence and sign language change in Nicaragua. In DeGraff 1999, 179–237.

Keil, Frank C. 1989. *Concepts, kinds, and cognitive development*. Cambridge, Mass.: MIT Press.

Kentridge, R. W., C. A. Heywood, and L. Weiskrantz. 1999. Attention without awareness in blindsight. *Proceedings of the Royal Society of London B* 266, 1805–1811.

Kinsbourne, Marcel. 1998. Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In José Luis Bermúdez, Anthony Marcel, and Naomi Eilan, eds., *The body and the self*, 205–223. Cambridge, Mass.: MIT Press.

Kintsch, Walter. 1974. *The representation of meaning in memory*. New York: John Wiley.

Kipper, Karin, and Martha Palmer. 2000. Representation of actions as an interlingua. In *Proceedings of the Third Workshop on Applied Interlinguas, held in conjunction with ANLP-NAACL*, Seattle.

Klein, Wolfgang, and Clive Perdue. 1997. The Basic Variety, or: Couldn't language be much simpler? *Second Language Research* 13, 301–347.

Klima, Edward S., and Ursula Bellugi. 1979. *The signs of language*. Cambridge, Mass.: Harvard University Press.

Koch, Christof. 2004. *The quest for consciousness*. Englewood, Colo.: Roberts.

Kohlberg, Lawrence. 1981–84. *The philosophy of moral development*. 2 vols. New York: Harper and Row.

Köhler, Wolfgang. 1927. *The mentality of apes*. London: Kegan Paul, Trench, Trubner and Co.

Kosslyn, Stephen. 1996. *Image and brain: The resolution of the imagery debate*. Cambridge, Mass.: MIT Press.

Lackner, James, and Paul Dizio. 2000. Aspects of body self-calibration. *Trends in Cognitive Sciences* 4, 279–288.

Ladd, D. Robert. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

Lakoff, George. 1970. *Irregularity in syntax*. New York: Holt, Rinehart and Winston.

Lakoff, George. 1971. On Generative Semantics. In Danny Steinberg and Leon Jakobovits, eds., *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*, 232–296. New York: Cambridge University Press.

Lakoff, George. 1987. *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Lakoff, George. 2002. *Moral politics: How liberals and conservatives think*. Chicago: University of Chicago Press.

Lakoff, George, and Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Lamb, Sydney. 1966. *Outline of Stratificational Grammar*. Washington, D.C.: Georgetown University Press.

Lamme, Victor A. F. 2003. Why visual attention and awareness are different. *Trends in Cognitive Sciences* 7, 12–18.

Landau, Barbara. 1996. Multiple geometric representations of objects in languages and language learners. In Bloom et al. 1996, 317–363.

Langacker, Ronald. 1987. *Foundations of Cognitive Grammar*, vol. 1. Stanford, Calif.: Stanford University Press.

Larson, Richard, and Gabriel Segal. 1995. *Knowledge of meaning: An introduction to semantic theory*. Cambridge, Mass.: MIT Press.

Lasersohn, Peter. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy* 28, 643–686.

Lenneberg, Eric. 1967. *Biological foundations of language.* New York: Wiley.

Lerdahl, Fred, and Ray Jackendoff. 1983. *A generative theory of tonal music*. Cambridge, Mass.: MIT Press.

Levelt, Willem J. M. 1989. *Speaking*. Cambridge, Mass.: MIT Press.

Levelt, Willem J. M. 1999. Producing spoken language: A blueprint of the speaker. In Colin M. Brown and Peter Hagoort, eds., *The neurocognition of language*, 83–122. Oxford: Oxford University Press.

Levin, Beth, and Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *Cognition* 41, 123–151.

Levin, Beth, and Malka Rappaport Hovav. 1995. *Unaccusativity*. Cambridge, Mass.: MIT Press.

Levinson, Stephen C. 2003. *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Levison, Libby, and Norman I. Badler. 1994. How animated agents perform tasks: Connecting planning and manipulations through object-specific reasoning. Presented at the AAAI Spring Symposium: Toward Physical Interaction and Manipulation, March 1994. Philadelphia: University of Pennsylvania, Department of Computer and Information Science.

Lewis, David. 1972. General semantics. In Donald Davidson and Gilbert Harman, eds., *Semantics of natural language*, 169–218. Dordrecht: Reidel.

Liberman, Mark, and Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8, 249–336.

Libet, Benjamin, C. A. Gleason, E. W. Wright, and D. K. Pearl. 1983. Time of conscious intention to act in relation to onset of cerebral activities (readiness potential): The unconscious initiation of a freely voluntary act. *Brain* 106, 623–642.

Lindsay, Peter H., and Donald A. Norman. 1977. *Human information processing: An introduction to psychology*. New York: Academic Press.

Linsky, Leonard, ed. 1971. *Reference and modality*. London: Oxford University Press.

Litman, Diane J., and James F. Allen. 1990. Discourse processing and commonsense plans. In Cohen, Morgan, and Pollack 1990, 365–388.

Mack, Arien, and Irvin Rock. 1998. *Inattentional blindness*. Cambridge, Mass.: MIT Press.

MacLennan, Bruce. 1997. The elements of consciousness and their neurodynamical correlates. In Shear 1997, 249–266.

Macnamara, John. 1991. The development of moral reasoning and the foundations of geometry. *Journal for the Theory of Social Behaviour* 21, 125–150.

Mahlmann, Matthias. 2003. Law and force: 20th century radical legal philosophy, post-modernism and the foundations of law. *Res Publica* 9, 19–37.

Mandler, George. 1993. Approaches to a psychology of value. In Hechter, Nadel, and Michod 1993, 229–258.

Marcus, Gary. 2001. *The algebraic mind*. Cambridge, Mass.: MIT Press.

Marr, David. 1982. *Vision*. San Francisco: Freeman.

Marr, David, and Lucia Vaina. 1982. Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London B* 214, 501–524.

McCarthy, John, and Patrick Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Bernard Meltzer and Donald Michie, eds., *Machine intelligence*, 4:463–502. New York: Harper and Row.

McCawley, James D. 1968. Lexical insertion in a transformational grammar without deep structure. In Bill Darden, Charles-James N. Bailey, and Alice Davison, eds., *Papers from the fourth regional meeting of the Chicago Linguistic Society*, 71–80. Chicago: University of Chicago, Chicago Linguistic Society.

McKay, Ryan, Robyn Langdon, and Max Coltheart. 2005. "Sleights of mind": Delusions, defences, and self-deception. *Cognitive Neuropsychiatry* 10, 305–326.

Metzinger, Thomas, ed. 2000. *Neural correlates of consciousness.* Cambridge, Mass.: MIT Press.

Mikhail, John. Forthcoming. *Rawls' linguistic analogy*. Cambridge: Cambridge University Press.

Miller, George A. 1956. The magical number seven plus or minus two: Some limits in our capacity for processing information. *Psychological Review* 63, 81–97.

Miller, George A., Eugene Galanter, and Karl H. Pribram. 1960. *Plans and the structure of behavior*. New York: Holt, Rinehart and Winston.

Miller, George A., and Philip Johnson-Laird. 1976. *Language and perception.* Cambridge, Mass.: Harvard University Press.

Millikan, Ruth. 1984. *Language, thought, and other biological categories.* Cambridge, Mass.: MIT Press.

Minsky, Marvin. 1968. Matter, mind, and models. In Marvin Minsky, ed., *Semantic information processing*, 425–432. Cambridge, Mass.: MIT Press.

Minsky, Marvin. 1975. A framework for representing knowledge. In Patrick H. Winston, ed., *The psychology of computer vision*, 211–277. New York: McGraw-Hill.

Minsky, Marvin. 1986. *The society of mind.* New York: Simon and Schuster.

Murphy, Gregory. 1996. On metaphoric representation. *Cognition* 60, 173–204.

Myung, Jong-Yoon, Sheila E. Blumstein, and Julie C. Sedivy. 2006. Playing on the typewriter, typing on the piano: Manipulation knowledge of objects. *Cognition* 98, 223–243.

Nelissen, Koen, Giuseppe Luppino, Wim Vanduffel, Giacomo Rizzolatti, and Guy A. Orban. 2005. Observing others: Multiple action representation in the frontal lobe. *Science* 310, 332–336.

Nemeroff, Carol, and Paul Rozin. 2000. The makings of the magical mind: The nature and function of sympathetic magical thinking. In Karl Rosengren, Carl Johnson, and Paul Harris, eds., *Imagining the impossible: Magical, scientific, and religious thinking in children*, 1–34. Cambridge: Cambridge University Press.

Newell, Allen, and Herbert A. Simon. 1972. *Human problem solving*. Englewood Cliffs, N.J.: Prentice-Hall.

Newmeyer, Frederick J. 1980. *Linguistic theory in America: The first quarter-century of transformational-generative grammar*. New York: Academic Press.

Newmeyer, Frederick J. 1998. On the supposed 'counterfunctionality' of Universal Grammar: Some evolutionary implications. In James Hurford, Michael Studdert-Kennedy, and Chris Knight, eds., *Approaches to the evolution of language,* 305–319. Cambridge: Cambridge University Press.

Newport, Elissa. 1990. Maturational constraints on language learning. *Cognitive Science* 14, 11–28.

Nisbett, Richard E., and Thomas D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84, 231–259.

Nowak, Martin, Joshua Plotkin, and Vincent A. A. Jansen. 2000. The evolution of syntactic communication. *Nature* 404, 495–498.

Nunberg, Geoffrey. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy* 3, 143–184.

Osaka, Naoyuki, ed. 2003. *Neural basis of consciousness*. Amsterdam: John Benjamins.

Partee, Barbara, ed. 1976. *Montague Grammar*. New York: Academic Press.

Parvizi, Josef, and Antonio Damasio. 2001. Consciousness and the brainstem. In Dehaene 2001, 135–160.

Perlmutter, David M., ed. 1983. *Studies in Relational Grammar 1*. Chicago: University of Chicago Press.

Pesetsky, David. 1995. *Zero syntax*. Cambridge, Mass.: MIT Press.

Phillips, Ann T., and Henry M. Wellman. 2006. Infants' understanding of object-directed action. *Cognition* 98, 137–155.

Piaget, Jean. 1932. *The moral judgment of the child*. Trans. Marjorie Gabain. New York: Free Press.

Pierrehumbert, Janet. 1980. The phonology and phonetics of English intonation. Doctoral dissertation, MIT.

Piñango, Maria Mercedes. 2000. Canonicity in Broca's sentence comprehension: The case of psychological verbs. In Yosef Grodzinsky, Lewis Shapiro, and David Swinney, eds., *Language and the brain: Representation and processing*, 330–350. San Diego, Calif.: Academic Press.

Piñango, Maria Mercedes, and Edgar Zurif. 2001. Semantic combinatorial operations in aphasic comprehension: Implications for the cortical localization of language. *Brain and Language* 79, 297–308.

Piñango, Maria Mercedes, Edgar Zurif, and Ray Jackendoff. 1999. Real-time processing implications of enriched composition at the syntax-semantics interface. *Journal of Psycholinguistic Research* 28, 395–414.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, Mass.: MIT Press.

Pinker, Steven. 1994. *The language instinct*. New York: HarperCollins.

Pinker, Steven. 1997. *How the mind works*. New York: W. W. Norton.

Pinker, Steven. 1999. *Words and rules*. New York: Basic Books.

Pinker, Steven. 2002. *The blank slate: The modern denial of human nature*. New York: Viking.

Pinker, Steven, and Paul Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13, 707–726.

Pinker, Steven, and Ray Jackendoff. 2005. The faculty of language: What's special about it? *Cognition* 95, 201–236.

Polanyi, Michael. 1958. *Personal knowledge*. Chicago: University of Chicago Press.

Pollack, Martha E. 1990. Plans as complex mental attitudes. In Cohen, Morgan, and Pollack 1990, 77–103.

Pollard, Carl, and Ivan Sag. 1987. *Information-based syntax and semantics*. Stanford, Calif.: CSLI Publications.

Pollard, Carl, and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Popper, Karl, and John Eccles. 1977. *The self and its brain*. New York: Springer International.

Posner, Michael. 1994. Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences* 91, 7398–7403.

Postal, Paul M. 1970. On the surface verb "remind." *Linguistic Inquiry* 1, 37–120.

Postal, Paul M. 1971. *Crossover phenomena*. New York: Holt, Rinehart and Winston.

Povinelli, Daniel J. 2000. *Folk physics for apes*. Oxford: Oxford University Press.

Premack, David. 1976. *Intelligence in ape and man*. Hillsdale, N.J.: Erlbaum.

Premack, David, and Ann James Premack. 1994. Moral belief: Form versus content. In Lawrence A. Hirschfeld and Susan Gelman, eds., *Mapping the mind: Domain specificity in cognition and culture*, 149–168. Cambridge: Cambridge University Press.

Premack, David, and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 515–526.

Prince, Alan, and Paul Smolensky. 1993. Optimality Theory: Constraint interaction in generative grammar. Technical report, Rutgers University and University of Colorado at Boulder.

Prinz, Jesse J. Forthcoming. The intermediate-level theory of consciousness. In Susan Schneider and Max Velmans, eds., *Blackwell companion to consciousness*. Oxford: Blackwell.

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, Mass.: MIT Press.

Putnam, Hilary. 1975. The meaning of 'meaning'. In Keith Gunderson, ed., *Language, mind, and knowledge*, 131–193. Minneapolis: University of Minnesota Press.

Pylyshyn, Zenon. 1984. *Computation and cognition*. Cambridge, Mass.: MIT Press.

Pylyshyn, Zenon. 2000. Situating vision in the world. *Trends in Cognitive Sciences* 4, 197–207.

Quine, W. V. O. 1956. Quantifiers and propositional attitudes. *Journal of Philosophy* 53, 177–187.

Ramachandran, V. S. 1995. Anosognosia in parietal lobe syndrome. *Consciousness and Cognition* 4, 22–51.

Rawls, John. 1971. *A theory of justice*. Cambridge, Mass.: Harvard University Press.

Revonsuo, Antti. 2000. Prospects for a scientific research program on consciousness. In Metzinger 2000, 57–75.

Rey, Georges. 2006. The intentional inexistence of language—but not cars. In Robert J. Stainton, ed., *Contemporary debates in cognitive science*, 237–256. Oxford: Blackwell.

Rizzolatti, Giacomo, L. Fadiga, V. Gallese, and L. Fogassi. 1996. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research* 3, 131–141.

Robinson, William S. 1997. The hardness of the Hard Problem. In Shear 1997, 149–161.

Rosenthal, David. 1986. Two concepts of consciousness. *Philosophical Studies* 94, 329–359.

Rumelhart, David, and James McClelland. 1986. On learning the past tense of English verbs. In James McClelland, David Rumelhart, and the PDP Research Group, *Parallel distributed processing*, 2:216–271. Cambridge, Mass.: MIT Press.

Russell, Bertrand. 1905. On denoting. *Mind* 14, 479–493.

Ryle, Gilbert. 1949. *The concept of mind.* Chicago: University of Chicago Press.

Sacerdoti, Earl D. 1977. *A structure for plans and behavior*. New York: American Elsevier.

Sadock, Jerrold. 1991. *Autolexical syntax*. Chicago: University of Chicago Press.

Sahlins, Marshall. 1976. *The use and abuse of biology: An anthropological critique of sociobiology*. Ann Arbor: University of Michigan Press.

Sandler, Wendy, Irit Meir, Carol Padden, and Mark Aronoff. 2005. The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences* 102, 2661–2665.

Saussure, Ferdinand de. 1915. *Cours de linguistique générale*. Ed. by Charles Bally and Albert Sechehaye. English translation: *Course in general linguistics*. New York: Philosophical Library, 1959.

Savage-Rumbaugh, Sue, Stuart Shanker, and Talbot Taylor. 1998. *Apes, language, and the human mind.* Oxford: Oxford University Press.

Schank, Roger, and Robert Abelson. 1975. *Scripts, plans, goals, and knowledge*. Hillsdale, N.J.: Erlbaum.

Scitovsky, Tibor. 1993. The meaning, nature, and sources of value in economics. In Hechter, Nadel, and Michod 1993, 93–105.

Searle, John. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417–424.

Searle, John. 1983. *Intentionality*. Cambridge: Cambridge University Press.

Searle, John. 1992. *The rediscovery of the mind.* Cambridge, Mass.: MIT Press.

Searle, John. 1995. *The construction of social reality.* New York: Free Press.

Sebanz, Natalie, Harold Bekkering, and Günther Knoblich. 2006. Joint action: Bodies and minds moving together. *Trends in Cognitive Sciences* 10, 70–76.

Seidenberg, Mark S., and Laura Petitto. 1978. Signing behavior in apes: A critical review. *Cognition* 17, 177–215.

Shapiro, Stuart C., and Haythem O. Ismail. 2003. Anchoring in a grounded layered architecture with integrated reasoning. *Robotics and Autonomous Systems* 43, 97–108.

Shear, Jonathan, ed. 1997. *Explaining consciousness: The Hard Problem.* Cambridge, Mass.: MIT Press.

Sherif, Muzafer, and Carolyn W. Sherif. 1966. *Groups in harmony and tension.* New York: Octagon.

Shieber, Stuart. 1986. *An introduction to unification-based approaches to grammar.* Stanford, Calif.: CSLI Publications.

Shore, Bradd. 1996. *Culture in mind.* New York: Oxford University Press.

Simmons, R. F. 1973. Semantic networks: Their computation and use for understanding English sentences. In Roger Schank and Kenneth M. Colby, eds., *Computer models of thought and language*, 63–113. San Francisco: Freeman.

Singer, W., A. K. Engel, A. K. Kreiter, M. H. J. Munk, S. Neuenschwander, and P. R. Roelfsema. 1997. Neuronal assemblies: Necessity, signature, and detectability. *Trends in Cognitive Sciences* 1, 252–260.

Singer, Wolf. 2000. Phenomenal awareness and consciousness from a neurobiological perspective. In Metzinger 2000, 121–137.

Smuts, Barbara B. 1985. *Sex and friendship in baboons.* New York: Aldine.

Smuts, Barbara B., Dorothy L. Cheney, Robert M. Seyfarth, Richard W. Wrangham, and Thomas T. Struhsaker, eds. 1987. *Primate societies.* Chicago: University of Chicago Press.

Snare, Frank. 1972. The concept of property. *American Philosophical Quarterly* 9, 200–206.

Solan, Lawrence M. 2005. Private language, public laws: The central role of legislative intent in statutory interpretation. *Georgetown Law Journal* 93, 427–486.

Spelke, Elizabeth S. 2003. Core knowledge. In Nancy Kanwisher and John Duncan, eds., *Attention and performance*. Vol. 20, *Functional neuroimaging of visual cognition*. Oxford: Oxford University Press.

Spelke, Elizabeth S., Gary Katz, Susan Purcell, Sheryl Ehrlich, and Karen Breinlinger. 1994. Early knowledge of object motion: Continuity and inertia. *Cognition* 51, 131–176.

Sripada, Chandra Sekhar, and Stephen Stich. Forthcoming. A framework for the psychology of norms. In Peter Carruthers, Stephen Laurence, and Stephen Stich, eds., *The innate mind.* Vol. 2, *Culture and cognition.* New York: Oxford University Press.

Stapp, Henry P. 1997. The Hard Problem: A quantum approach. In Shear 1997, 197–215.

Stevens, Jeffrey R., and Marc D. Hauser. 2004. Why be nice? Psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences* 8, 60–65.

Stich, Stephen. 1983. *From folk psychology to cognitive science: The case against belief*. Cambridge, Mass.: MIT Press.

Stich, Stephen. 1993. Moral philosophy and mental representation. In Hechter, Nadel, and Michod 1993, 215–228.

Stone, Julius. 1968. *Human law and human justice.* Stanford, Calif.: Stanford University Press.

Swinney, David. 1979. Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18, 645–659.

Talmy, Leonard. 1983. How language structures space. In Herbert Pick and Linda Acredolo, eds., *Spatial orientation: Theory, research, and application*, 225–282. New York: Plenum. Revised version in Talmy 2000, 1:177–254.

Talmy, Leonard. 1988. Force dynamics in language and thought. *Cognitive Science* 12, 49–100. Revised version in Talmy 2000, 1:409–470.

Talmy, Leonard. 2000. *Toward a cognitive semantics.* 2 vols. Cambridge, Mass.: MIT Press.

Tanenhaus, Michael K., James M. Leiman, and Mark Seidenberg. 1979. Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior* 18, 427–440.

Tenny, Carol. 1994. *Aspectual roles and the syntax-semantics interface*. Dordrecht: Kluwer.

Terrace, Herbert. 1979. *Nim.* New York: Knopf.

Tetlock, Philip E. 2003. Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences* 7, 320–324.

Tomasello, Michael. 1998. *The new psychology of language*. Hillsdale, N.J.: Erlbaum.

Tomasello, Michael, ed. 2000. Primate cognition. Special issue, *Cognitive Science* 24.3.

Tomasello, Michael. 2003. *Constructing a language*. Cambridge, Mass.: Harvard University Press.

Tomasello, Michael, Josep Call, and Brian Hare. 2003. Chimpanzees understand psychological states—the question is which ones and to what extent. *Trends in Cognitive Sciences* 7, 153–156.

Tomasello, Michael, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28, 675–691.

Tooby, John, and Leda Cosmides. 1989. Evolutionary psychology and the generation of culture, part I. *Ethology and Sociobiology* 10, 29–49.

Tooby, John, and Leda Cosmides. 1992. The psychological foundations of culture. In Jerome Barkow, Leda Cosmides, and John Tooby, eds., *The adapted mind*, 19–136. New York: Oxford University Press.

Treisman, Ann. 1988. Features and objects: The fourteenth Bartlett Memorial Lecture. *Quarterly Journal of Experimental Psychology* 40A, 201–237.

Trivers, Robert L. 1971. The evolution of reciprocal altruism. *Quarterly Review of Biology* 46, 35–57.

Turiel, Elliott. 1983. *The development of social knowledge: Morality and convention*. Cambridge: Cambridge University Press.

Turillo, Carmelo Joseph, Robert Folger, James J. Lavelle, Elizabeth E. Umphress, and Julie O. Gee. 2002. Is virtue its own reward? Self-sacrificial decisions for the sake of fairness. *Organizational Behavior and Human Decision Processes* 89, 839–865.

Tversky, Barbara, Jeffrey M. Zacks, and Paul Lee. 2004. Events by hands and feet. *Spatial Cognition and Computation* 4, 5–14.

Ullman, Shimon. 1998. Three-dimensional object recognition based on the combination of views. *Cognition* 67, 21–44.

van der Zee, Emile, and Jon Slack, eds. 2003. *Representing direction in language and space*. Oxford: Oxford University Press.

van Schaik, Carel P., Marc Ancrenaz, Gwendolyn Borgen, Birute Galdika, Cheryl D. Knott, Ian Singleton, Akira Suzuki, Sri Suci Utani, and Michelle Merrill. 2003. Orangutan cultures and the evolution of material culture. *Science* 299, 102–105.

Van Valin, Robert, and Randy LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

Vandeloise, Claude. 1986. *L'espace en français*. Paris: Editions du Seuil.

Varela, Francisco J. 1997. Neurophenomenology: A methodological remedy for the Hard Problem. In Shear 1997, 337–357.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The embodied mind*. Cambridge, Mass.: MIT Press.

Verkuyl, Henk. 1993. *A theory of aspectuality: The interaction between temporal and atemporal structure*. Cambridge: Cambridge University Press.

von Wright, Georg Henrik. 1963. *Norm and action: A logical enquiry.* London: Routledge and Kegan Paul.

Watanabe, John M., and Barbara B. Smuts. 2004. Cooperation, commitment, and communication in the evolution of human sociality. In Robert W. Sussman and Audrey R. Chapman, eds., *The origins and nature of sociality*, 288–309. New York: Aldine de Gruyter.

Wegner, Daniel. 2002. *The illusion of conscious will*. Cambridge, Mass.: MIT Press.

Weinreich, Uriel. 1966. Explorations in semantic theory. In Thomas Sebeok, ed., *Current trends in linguistics*, 3:395–477. The Hague: Mouton. Reprinted in Uriel

Weinreich, *On semantics*, 99–201. Philadelphia: University of Pennsylvania Press, 1980.

Werner, Oswald, and Martin D. Topper. 1976. On the theoretical unity of ethnoscience lexicography and ethnoscience ethnographies. In Clea Rameh, ed., *Georgetown University Round Table on Languages and Linguistics 1976. Semantics: Theory and application*, 111–143. Washington, D.C.: Georgetown University Press.

Wexler, Kenneth, and Peter Culicover. 1980. *Formal principles of language acquisition*. Cambridge, Mass.: MIT Press.

Whiten, Andrew. 2002. The imitator's representation of the imitated: Ape and child. In Andrew N. Meltzoff and Wolfgang Prinz, eds., *The imitative mind: Development, evolution and brain bases*, 98–121. Cambridge: Cambridge University Press.

Whiten, Andrew, J. Goodall, W. C. McGrew, T. Nishida, V. Reynolds, Y. Sugiyama, G. E. G. Tutin, R. W. Wrangham, and C. Boesch. 2001. Charting cultural variation in chimpanzees. *Behavior* 138, 1481–1516.

Wierzbicka, Anna. 1987. *English speech act verbs: A semantic dictionary*. Sydney: Academic Press.

Williams, Edwin. 1994. Remarks on lexical knowledge. In Lila Gleitman and Barbara Landau, eds., *The acquisition of the lexicon*, 7–34. Cambridge, Mass.: MIT Press.

Wilson, Edward O. 1975. *Sociobiology: The new synthesis*. Cambridge, Mass.: Harvard University Press.

Yip, Moira. 1995. Tone in East Asian languages. In Goldsmith 1995, 476–494.

Yip, Moira, Joan Maling, and Ray Jackendoff. 1987. Case in tiers. *Language* 63, 217–250.

Zacks, Jeffrey M., and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological Review* 127, 3–21.

Zubizarreta, Maria Luisa. 1988. The lexical encoding of scope relations among arguments. Ms., University of Maryland, College Park.

Zurif, Edgar. 1990. Language and the brain. In Daniel N. Osherson and Howard Lasnik, eds., *An invitation to cognitive science*. Vol. 1, *Language*, 177–198. Cambridge, Mass.: MIT Press.

Zuriff, Gerald. 1998. Against metaphysical social constructionism in psychology. *Behavior and Philosophy* 26, 5–28.

Zwicky, Arnold. 1994. Dealing out meaning: Fundamentals of syntactic constructions. In *Proceedings of the 20th Annual Meeting of the Berkeley Linguistics Society*, 611–625. Berkeley: University of California, Berkeley Linguistics Society.

Index